Fine-grained Contrastive Learning for Relation Extraction

William Hogan **Jiacheng Li** **Jingbo Shang***
Department of Computer Science & Engineering
University of California, San Diego
{whogan,j9li,jshang}@ucsd.edu

2023. 2. 2 • ChongQing

—— EMNLP 2022

https://github.com/wphogan/finecl

Reported by JiaWei Cheng

Motivation

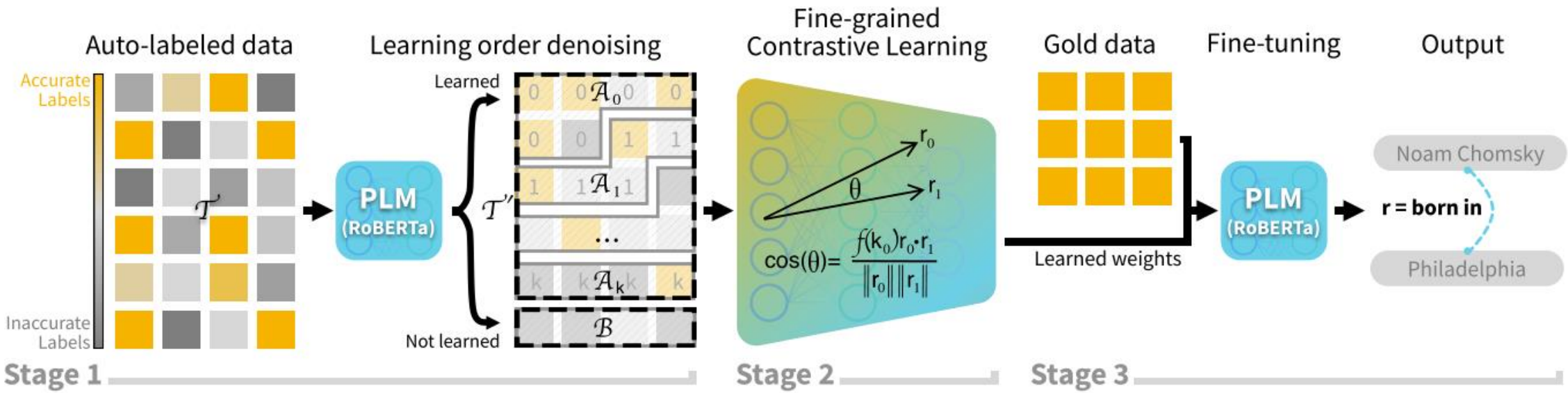
1. "*Noam Chomsky* was born in *Philadelphia*."
2. "*Noam Chomsky* gave a presentation in *Philadelphia*."
3. "Raised in the streets of *Philadelphia*, *Noam Chomsky*..."

[*Noam Chomsky*, *born in*, *Philadelphia*]

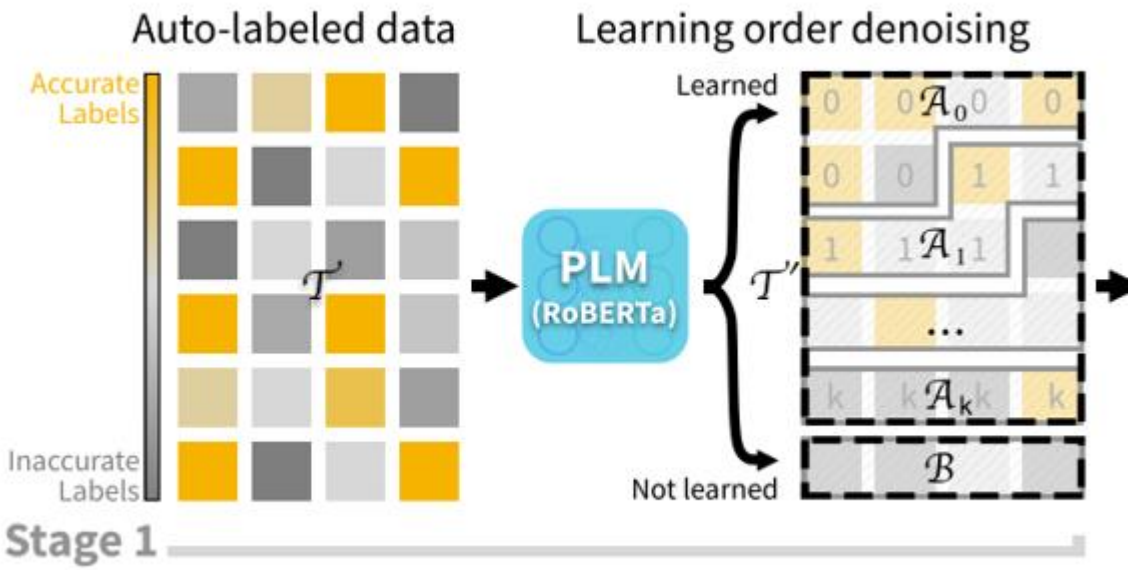
1.The authors argue that distantly supervised relation data contains a large number of mislabeled data

2.Conventional contrastive learning for RE does not account for differences in label accuracy—it treats all instances equally.

Overview


Auto-labeled data
Learning order denoising
Fine-grained Contrastive Learning
Gold data
Fine-tuning
Output
Accurate Labels
Inaccurate Labels
PLM (RoBERTa)
Learned
Not learned
Learned weights
PLM (RoBERTa)
Noam Chomsky
r = born in
Philadelphia
Stage 1
Stage 2
Stage 3

Method

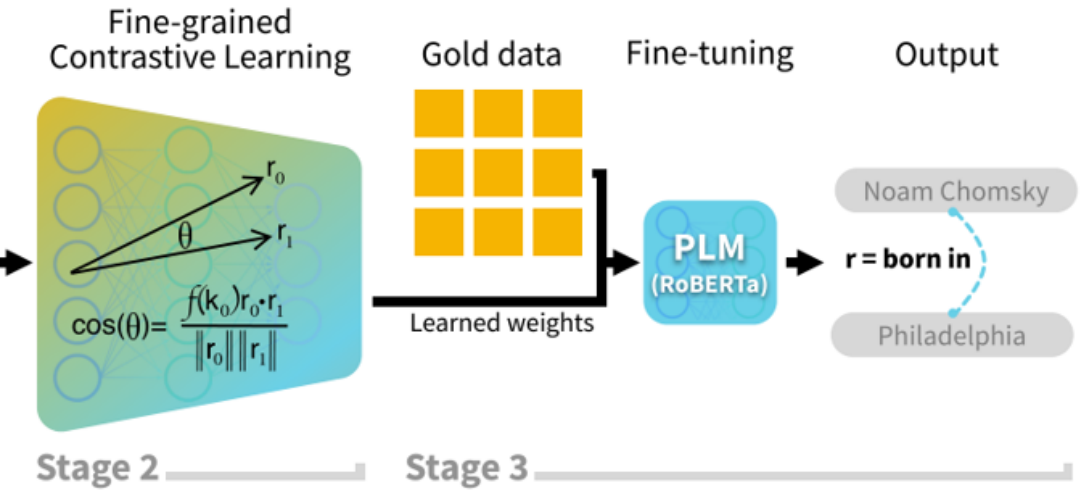


$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{N} y_{o,i} \cdot \log\left(p\left(y_{o,i}\right)\right) \tag{1}$$

$$\mathcal{A}_0 \bigcup \mathcal{A}_1 \ldots \bigcup \mathcal{A}_k = \mathcal{A} \tag{2}$$

$$\mathcal{A}_i \bigcap \mathcal{A}_j = \emptyset \text{ for all } i \neq j \tag{3}$$

Method



$$\mathbf{m}_{e_j} = \text{MeanPool}(\mathbf{h}_{n_{\text{start}}}, \ldots, \mathbf{h}_{n_{\text{end}}}), \quad (4)$$

$$\mathcal{L}_{\text{E}_\text{D}} = -\sum_{t^i_{jk} \in \mathcal{T}'} \log \frac{\exp(\cos(\mathbf{e}_{ij}, \mathbf{e}_{ik})/\tau)}{\sum_{l=1, l\neq j}^{|\mathcal{E}_i|} \exp(\cos(\mathbf{e}_{ij}, \mathbf{e}_{il})/\tau)}$$

$$\mathcal{L}_{\text{R}_\text{D}} = -\sum_{t_A, t_B \in \mathcal{T}'} f(k_A) \log \frac{\exp\left(\cos\left(\mathbf{r}_{t_A}, \mathbf{r}_{t_B}\right)/\tau\right)}{\mathcal{Z}},$$

$$\mathcal{Z} = \sum_{t_C \in \mathcal{T}'/\{t_A\}}^{N} f(k_C) \exp\left(\cos\left(\mathbf{r}_{t_A}, \mathbf{r}_{t_C}\right)/\tau\right) \quad (5)$$

$$f(k) = \alpha^{\frac{k_{\max} - k}{k_{\max} - k_{\min}}}, \quad (6)$$

$$\mathcal{L}_{FineCL} = \mathcal{L}_{E_D} + \mathcal{L}_{R_D} + \mathcal{L}_{MLM} \quad (7)$$

Experiments

	Base Lang. Model	Pre-train objective	R_D	E_D
BERT	BERT	MLM	×	×
RoBERTa	RoBERTa	MLM	×	×
MTB	BERT	DPS	✓	×
CP	BERT	CL + MLM	✓	×
$ERICA_{BERT}$	BERT	CL + MLM	✓	✓
$ERICA_{RoBERTa}$	RoBERTa	CL + MLM	✓	✓
WCL	BERT	WCL + MLM	✓	×
FineCL	RoBERTa	FineCL + MLM	✓	✓

Table 1: A comparison of RE pre-training methods highlighting the pre-training objective: Mask Language Modeling (MLM), Dot Product Similarity (DPS), Contrastive Learning (CL), Weighted Contrastive Learning (WCL), and Fine-grained Contrastive Learning (FineCL). R_D denotes the presence of relation discrimination in the loss function, and E_D denotes the presence of entity discrimination in the loss function.

Experiments

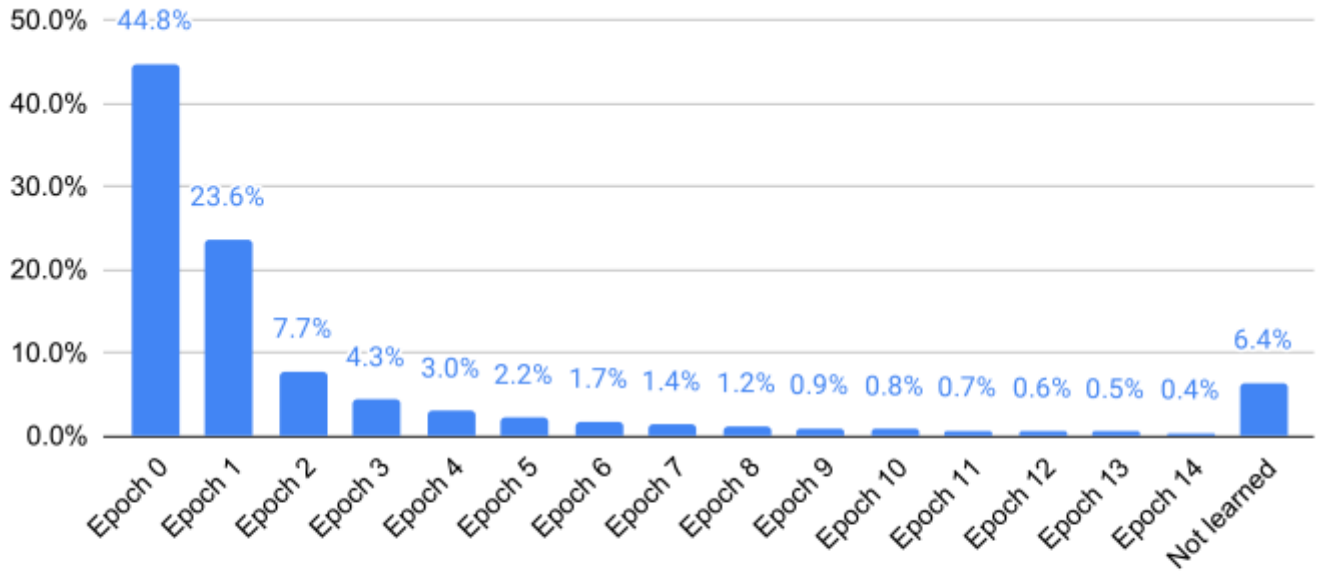


Figure 2: Percent of total training instances learned per epoch when recording *batch-based* learning order on distantly labeled data from DocRED.

Experiments

Size	1%		10%		100%	
Metrics	F1	IgF1	F1	IgF1	F1	IgF1
CNN*	-	-	-	-	42.3	40.3
BiLSTM*	-	-	-	-	51.1	50.3
HINBERT*	-	-	-	-	55.6	53.7
CorefBERT*	32.8	**31.2**	46.0	43.7	57.0	54.5
SpanBERT*	32.2	30.4	46.4	44.5	57.3	55.0
ERNIE*	26.7	25.5	46.7	44.2	56.6	54.2
MTB*	29.0	27.6	46.1	44.1	56.9	54.3
CP*	30.3	28.7	44.8	42.6	55.2	52.7
BERT	19.9	18.8	45.2	43.1	56.6	54.4
RoBERTa	29.6	27.9	47.6	45.7	58.2	55.9
$\text{ERICA}_{\text{BERT}}$	22.9	21.7	48.5	46.4	57.4	55.2
$\text{ERICA}_{\text{RoBERTa}}$	30.0	28.2	50.1	48.1	59.1	56.9
$\text{WCL}_{\text{RoBERTa}}$	22.3	20.8	49.4	47.5	58.5	56.2
FineCL	**33.2**	**31.2**	**50.3**	**48.3**	**59.5**	**57.1**

Table 3: F1-micro scores reported on the DocRED test set. IgF1 ignores performance on fact triples in the test set overlapping with triples in the train/dev sets. (* denotes performance as reported in (Qin et al., 2021); all other numbers are from our implementations).

Experiments

Metric	F1-macro	F1-macro-weighted
BERT	37.3	54.9
RoBERTa	39.6	56.9
$\text{ERICA}_{\text{BERT}}$	37.9	55.8
$\text{ERICA}_{\text{RoBERTa}}$	<u>40.1</u>	<u>57.8</u>
$\text{WCL}_{\text{RoBERTA}}$	39.9	57.2
FineCL	**40.7**	**58.2**

Table 4: F1-macro and F1-macro-weighted scores reported from the DocRED test set.

Experiments

Dataset	TACRED			SemEval		
Size	1%	10%	100%	1%	10%	100%
MTB*	35.7	58.8	68.2	44.2	79.2	88.2
CP*	37.1	60.6	68.1	40.3	80.0	88.5
BERT	22.2	53.5	63.7	41.0	76.5	87.8
RoBERTa	27.3	61.1	69.3	43.6	77.7	87.5
$\text{ERICA}_{\text{BERT}}$	34.9	56.0	64.9	46.4	79.8	88.1
$\text{ERICA}_{\text{RoBERTa}}$	41.1	61.7	69.5	50.3	80.9	88.4
$\text{WCL}_{\text{RoBERTA}}$	37.6	61.3	69.7	47.0	80.0	88.3
FineCL	**43.7**	**62.7**	**70.3**	**51.2**	**81.0**	**88.7**

Table 5: F1-micro scores reported from the TACRED and SemEval test sets (* denotes performance as reported in (Qin et al., 2021); all other numbers are from our implementations).

Experiments

Epochs of learning order data	% Learned	F1	IgF1
Baseline	N/A	58.7	56.5
1 Epoch	45	58.6	56.4
3 Epochs	76	58.6	56.3
5 Epochs	83	58.7	56.5
10 Epochs	92	58.8	56.6
15 Epochs	94	**59.0**	**56.7**

Table 6: Ablation experiment results on the DocRED test set with pre-trained models that use learning order data obtained with various training durations. Percent learned refers to the percent of training instances learned in the set of learned instances ($\mathcal{A}$). "Baseline" is a pre-trained model that does not leverage learning order (i.e., all instances are weighted equally during pre-training).

Experiments

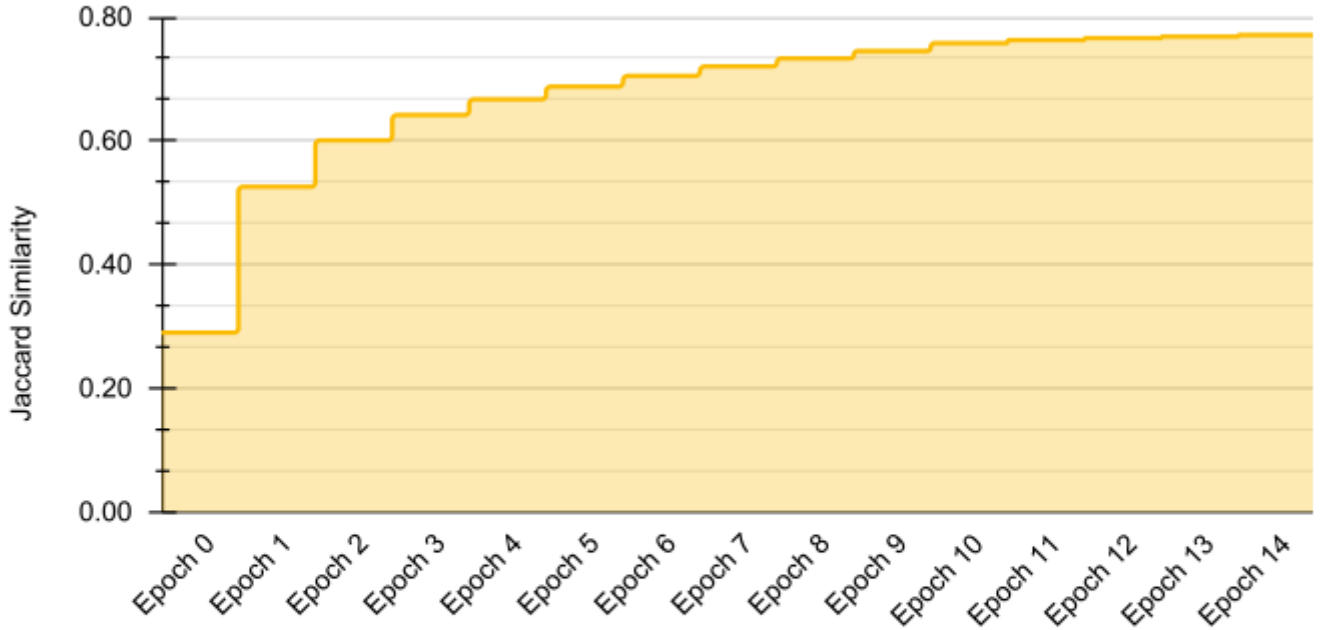


Figure 3: Cumulative Jaccard Similarity between sets of learned instances by epoch from RoBERTa and SSAN using distantly labeled training data from DocRED.

Thanks!