Continual Contrastive Finetuning Improves Low-Resource Relation Extraction

Wenxuan Zhou[1], Sheng Zhang[2], Tristan Naumann[2], Muhao Chen[1], Hoifung Poon[2]
University of Southern California[1], Microsoft Research[2]

Motivation

Manual annotation for relation extraction is expensive

~$12 per passage (~200 words) in the general domain (Tan et al., 21)

~$138 per PubMed abstract (~235 words, Luo et al., 22)

Unlabeled data is abundant and easy to acquire

WIKIPEDIA
The Free Encyclopedia
6.6M articles

PubMed
35M citations and abstracts

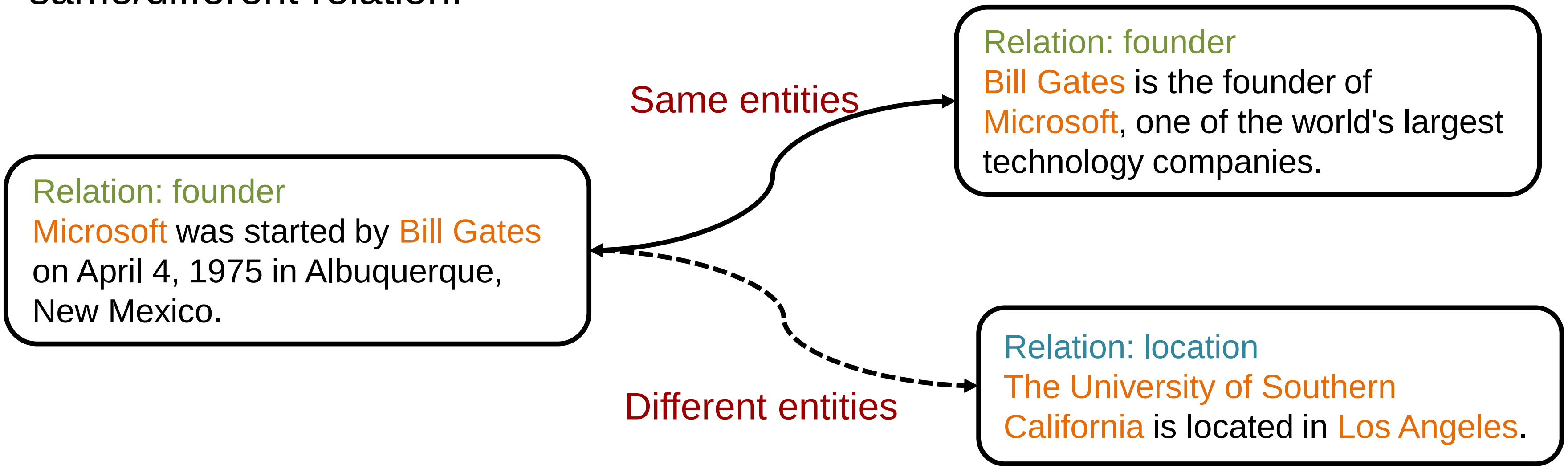
bioRxiv
180K research papers

How to use unlabeled data to improve relation extraction?

Matching the Blanks (MTB)

Harris's distributional hypothesis to relations (Soares et al., 19):

Context linking the same/different entities is more likely to express the same/different relation.



MTB-based Pretraining (Soares et al., 19)

Positive instances: instances with the same entities

Negative instances: instances with different entities

Goal: make embedding of positive/negative instance pairs similar/dissimilar.

(Bill Gates, Microsoft)

was started by
on April 4, 1975 in Albuquerque,
New Mexico.

Positive instances
Increase similarity

(Bill Gates, Microsoft)

is the founder of
one of the world's largest
technology companies.

Negative instances
Reduce similarity

(USC, Los Angeles)

is located in

Continual Contrastive Finetuning

Previous work

Pretraining
Pairwise similarity (MTB)

Finetuning
Classification

Mismatch in objectives

Continual contrastive finetuning

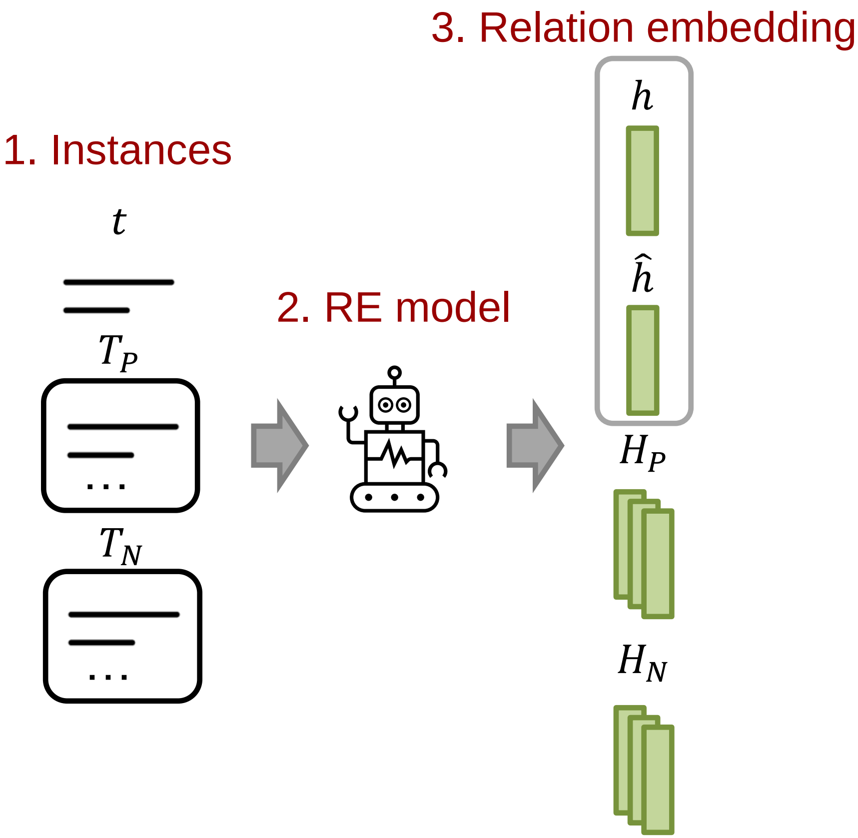
Pretraining
Contrastive pretraining

Finetuning
Multi-cluster contrastive finetuning

Inference
Similarity-based kNN

Consistent objectives

Contrastive Pretraining



H': all relation embedding of all instances except for t.
τ: temperature

MTB-based contrastive loss:
Make embedding between t and its positive/negative instances to be similar/dissimilar.

$$L_{\text{mtb}} = -\frac{1}{|P|} \sum_{h_1 \in P} \log \left(\frac{e^{cos(h,h_1)/\tau}}{Z} \right)$$

$$Z = e^{cos(h,h_1)/\tau} + \sum_{h_2 \in N} e^{cos(h,h_2)/\tau}$$

Self-supervised contrastive loss:
Make the similarity between two embeddings of t to be larger than t and other instances.

$$L_{\text{self}} = -\log \left(\frac{e^{cos(h,\hat{h})/\tau}}{Z} \right)$$

$$Z = \sum_{h_2 \in H'} e^{cos(h,h_2)/\tau}$$

Masked language modeling loss L_{mlm}.

Joint training:

$$L_{\text{pretrain}} = L_{\text{mtb}} + L_{\text{self}} + L_{\text{mlm}}$$

Distributional Gap

Make embedding from the same/different classes similar/dissimilar

Classic finetuning objectives:

Softmax classifier with cross-entropy (CE), supervised contrastive loss (SupCon)

Distributional gap:

- CE and SupCon are minimized when representations form a single cluster for a class. (Graf et al., 2021)
- Representations from MTB-based pretraining may form multiple clusters for a class.

The headquarter of Honda Corp is located in Japan.

Mount Fuji is the highest mountain in Japan.

Pretraining
Negative instances!

Finetuning
Same relation of "*location*"

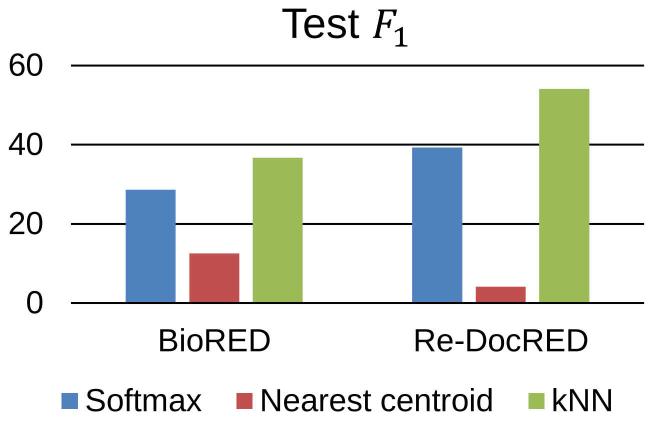
Distributional Gap (Cont.)

Probing analysis:

Given an MTB-based pretrained RE model, fix the model parameters and fit different classifiers on top of it.

Classifiers:

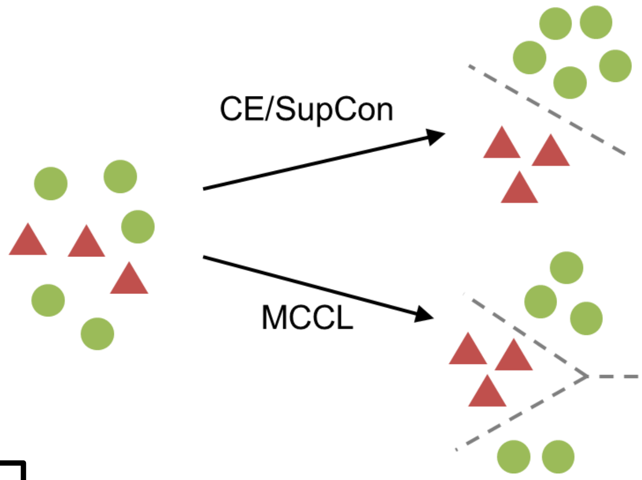
- Single-cluster: softmax classifier, nearest centroid classifier
- Multi-cluster: kNN



KNN greatly outperforms both softmax and nearest centroid ⇒ MTB-based pretraining generates multi-cluster representations.

Multi-cluster Contrastive Learning

Goal: learn multiple clusters for a class.



Step 1: weighted class score

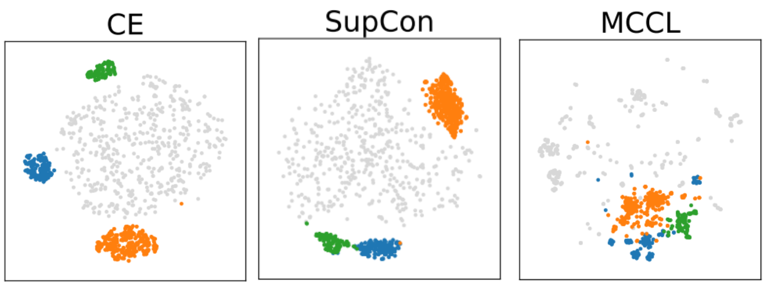
For instance t_i and relation r, denote instances of relation r as T_r, the score for t_i and r is

$$score_{(t_i,r)} = \sum_{t_j \in T_r \setminus \{t_i\}} weight_{(t_i,t_j)} \cdot cos(h_i, h_j)$$

Instance weight, larger for more similar instances. Calculate by softmax.

Step 2: Classification

Make the score of the true class larger than others with cross-entropy loss.



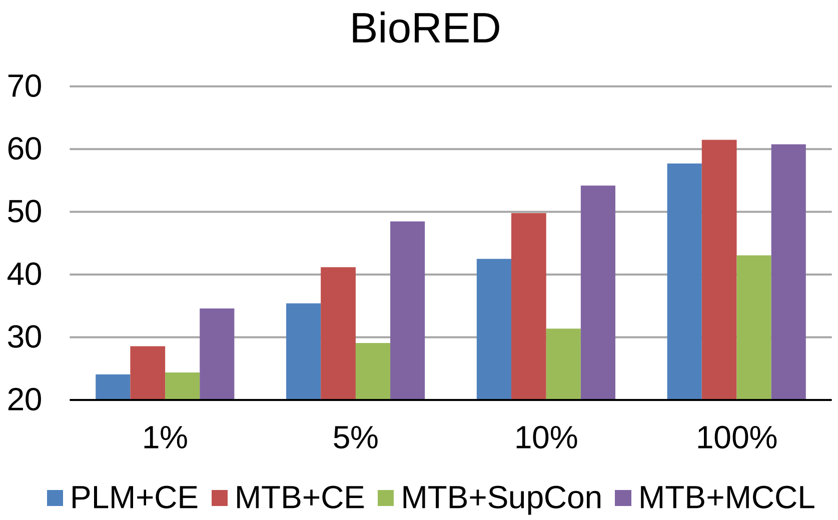
t-SNE visualization of relation embedding

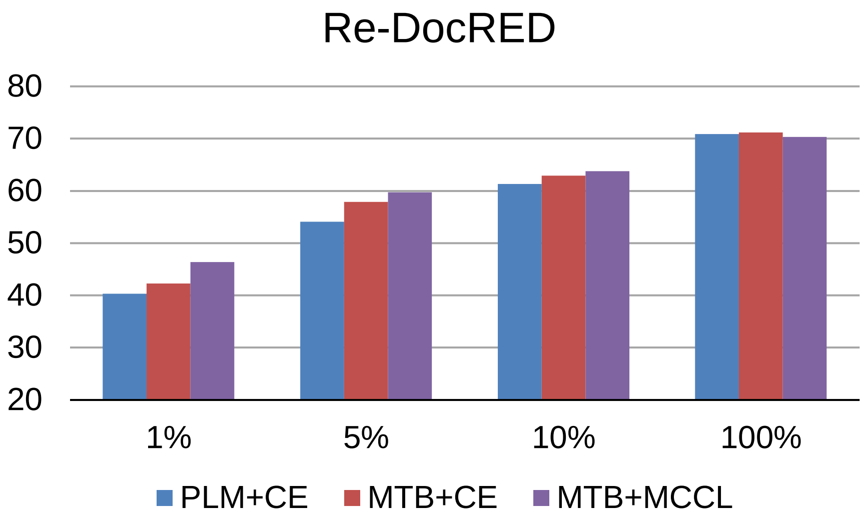
Experiments: Main Results

Datasets: BioRED, Re-DocRED (multi-label)

Models: PubmedBERT for BioRED, BERT for Re-DocRED.

Evaluation metric: F1

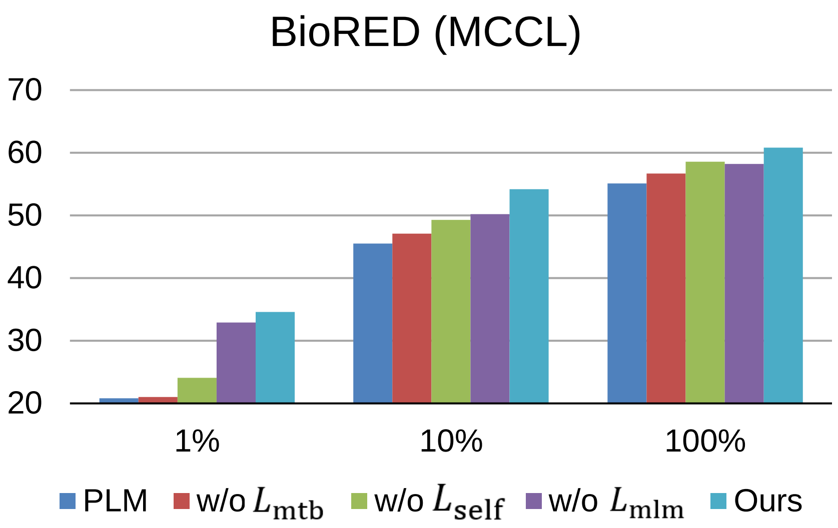




1. MCCL consistently outperforms CE and SupCon in low-resource settings.
2. MTB+CE outperforms PLM+CE, showing that MTB pretraining is effective.

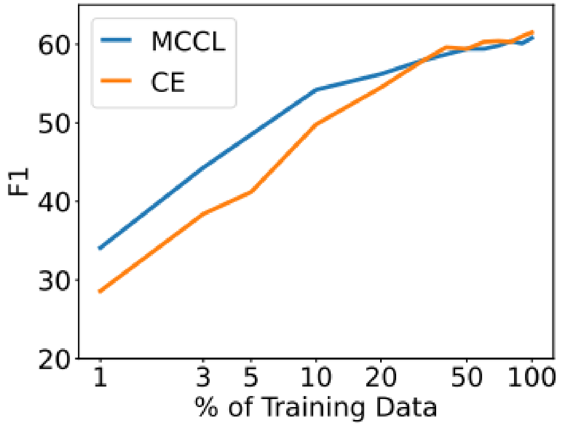
Experiments: Additional Analysis

Ablation Study



1. All the pretraining objectives are effective.
2. Removing MTB leads to the largest drop ⇒ MTB is critical for low-resource RE

F_1 w.r.t. different % of data



1. MCCL consistently outperforms CE when $< 20\%$ of data (~80 abstracts) are used.
2. MCCL performs similarly to CE with abundant training data.

Conclusion

1. We propose to pretrain the PLMs based on our improved MTB objective and show that it greatly improves PLM performance in low-resource document-level RE.
2. We bridge the gap of learning objectives between RE pretraining and finetuning with continual contrastive finetuning and kNN-based inference, helping the RE model leverage pretraining knowledge.
3. We design a multi-cluster contrastive learning objective, allowing one relation to form multiple different clusters, thus further reducing the distributional gap between pretraining and finetuning.