HiCLRE: A Hierarchical Contrastive Learning Framework for Distantly Supervised Relation Extraction

Dongyang Li[1,2] [*] **Taolin Zhang**[3,4][*] **Nan Hu**[1], **Chengyu Wang**[3], **Xiaofeng He**[1,5][†]

[1] School of Computer Science and Technology, East China Normal University
[2] Shanghai Key Laboratory of Trsustworthy Computing [3] Alibaba Group
[4] School of Software Engineering, East China Normal University
[5] Shanghai Research Institute for Intelligent Autonomous Systems

`dongyangli0612@gmail.com, zhangtl0519@gmail.com`
`hunan.vinny1997@gmail.com, chengyu.wcy@alibaba-inc.com`
`hexf@cs.ecnu.edu.cn`

Abstract

Distant supervision assumes that any sentence containing the same entity pairs reflects identical relationships. Previous works of distantly supervised relation extraction (DSRE) task generally focus on sentence-level or bag-level denoising techniques independently, neglecting the explicit interaction with cross levels. In this paper, we propose a **Hi**erarchical **C**ontrastive **L**earning Framework for Distantly Supervised **R**elation **E**xtraction (HiCLRE) to reduce noisy sentences, which integrate the global structural information and local fine-grained interaction. Specifically, we propose a three-level hierarchical learning framework to interact with cross levels, generating the de-noising context-aware representations via adapting the existing multi-head self-attention, named Multi-Granularity Recontextualization. Meanwhile, pseudo positive samples are also provided in the specific level for contrastive learning via a dynamic gradient-based data augmentation strategy, named Dynamic Gradient Adversarial Perturbation. Experiments demonstrate that Hi-CLRE significantly outperforms strong baselines in various mainstream DSRE datasets.[1]

1 Introduction

Relation extraction (RE) can draw relations of two entities from unstructured text. It can be widely used in natural language processing applications such as knowledge graph construction (Khatib et al., 2020; Tang et al., 2020) and question answering (Wang and Jiang, 2019; Liu et al., 2020; Saxena et al., 2020). Existing RE works (Wei et al., 2020; Alt et al., 2020; Veyseh et al., 2020) rely on a large-scale annotated dataset, which is time-consuming and labor-intensive. DSRE (Mintz et al.,

---

[*]D. Li and T. Zhang contributed equally to this work.
[†]Corresponding author.
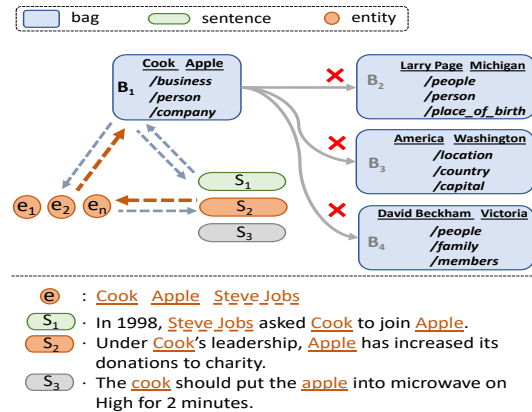[1]The source code and data can be available at https://github.com/MatNLP/HiCLRE

Figure 1: Example of semantic relationships in specific levels and cross levels. The red cross means the semantic difference of two bag-level relations and the dotted arrow indicates the semantic overlapping of cross levels. (Best viewed in color).

2009) attempts to address this issue via automatically generating training text samples. Obviously, this assumption introduces noisy data and may hurt the performance. Hence, multi-instance learning (MIL) (Zeng et al., 2015) is further proposed to assign a bag containing *"at least one"* correct sentence of relation triple.

The previous approaches of DSRE tackle the task at different granularities (i.e. sentence-level and bag-level). (1) Sentence-level. These works (Wu et al., 2019; Li et al., 2019) focus on finding the ground-truth relational labels from the internal semantics of the input sentences. (2) Bag-level. Although these works (Su et al., 2018; Beltagy et al., 2019; Chen et al., 2021a; Christopoulou et al., 2021) consider the information of sentence-level and bag-level simultaneously, but they ignore the explicit cross-level interactions, which contain plenty of knowledge to further boost the DSRE task performance. As shown in Figure 1, the rich semantic information of bag-level and sentence-

level are provided for the "Cook" and "Apple" in the entity level. For example, "Steve Jobs" in s_1 is also the co-founder of "Apple" company and the label of bag-level is the "*/business/person/company*" exactly shows the relation of this entity pair. Meanwhile, the huge semantic difference exists in a specific level such as the "*/business/person/company*" and "*/location/country/capital*" in bag level.

To overcome the challenges mentioned above, we propose a **Hi**erarchical **C**ontrastive **L**earning framework for distantly supervised **R**elation **E**xtraction (HiCLRE), which facilities semantic interactions within a specific level and cross levels:

(1) Multi-Granularity Recontextualization: To capture the cross-level structural information, we adapt the multi-head self-attention mechanism into three-level granularities, including entity-level, sentence-level and bag-level. We align the context-aware feature of each layer with the input of attention mechanism respectively. The refined representations as recontextualized interaction semantics are picked out for the corresponding level via the attention scores aggregated by the other two levels.

(2) Dynamic Gradient Adversarial Perturbation: To obtain the more accurate specific-level representations, we employ gradient-based contrastive learning (Hadsell et al., 2006; van den Oord et al., 2018) to pull the information of constructed pseudo positive samples and push the difference of negative samples. Concretely, we calculate the dynamic perturbation from two aspects, including the normalized gradient of task loss and the temporal weighted memories similarity between the last and current epoch.

To verify the effectiveness of HiCLRE, we evaluate our model on three mainstream DSRE datasets, including NYT10 (Riedel et al., 2010), GDS (Jat et al., 2017), and KBP (Ling and Weld, 2012). The experimental results show that HiCLRE significantly outperforms the state-of-the-art baselines' performance, achieving a 2.2% relative AUC increase and improving the P@M score from 77.2% to 78.2%. Furthermore, the ablation study shows the individual contributions of each module.

Accordingly, the major contributions of this paper are summarized as follows:

- We propose a hierarchical contrastive learning framework for DSRE task (HiCLRE), which fully utilizes the semantic interaction within the specific level and cross levels, reducing the influence of noisy data.

- The multi-granularity recontextualization is proposed to enhance the cross-level interaction and the dynamic gradient adversarial perturbation learns better representations within three specific levels.

- Extensive experiments show that our model outperforms the strong baseline over DSRE datasets and detailed analysis demonstrates the modules are also effective.

2 Related Work

2.1 Distantly Supervised Relation Extraction

Recently, these works are divided into two categories. (1) Human-designed Feature. (Yao et al., 2011) propose three types of LDA (i.e. Rel-LDA, Rel-LDA1, and Type-LDA) to cluster the similar triples together. MIML (Hoffmann et al., 2011; Surdeanu et al., 2012) and MIL (Zeng et al., 2015) attempt to relax the limitation of distantly supervision assumption to tackle the data generation problem. (2) Neural Networks Representation. These models automatically generate the feature representation via end-to-end learning to reduce manual intervention. (Qin et al., 2018) introduce a generative adversarial training framework that provides a cleaned dataset for RE task. (Ye and Ling, 2019) consider both inter-bag and intra-bag attention to handle the noise at sentence-level and bag-level independently. SENT (Ma et al., 2021) is a sentence-level framework to generate efficient training samples by negative training to filter the noisy data. These works generally use the partial levels' information independently to explore the relational semantics.

2.2 Contrastive Learning

Loss Function NCE (Gutmann and Hyvärinen, 2010) learns a classifier to distinguish the clean and noisy examples with the probability density function. InfoNCE (van den Oord et al., 2018) integrates the mutual information into the NCE, which can maximize similarity and minimize the difference.

Data Augmentation These works can be generally divided into three categories. (1) Data augmentation by simple text processing. EDA (Wei and Zou, 2019) proposes synonyms replace, randomly insert and randomly delete operations. CIL (Chen et al., 2021a) utilizes TF-IDF scores to insert/substitute some unimportant words to/in instance to construct positive samples. (2) Data augmentation by
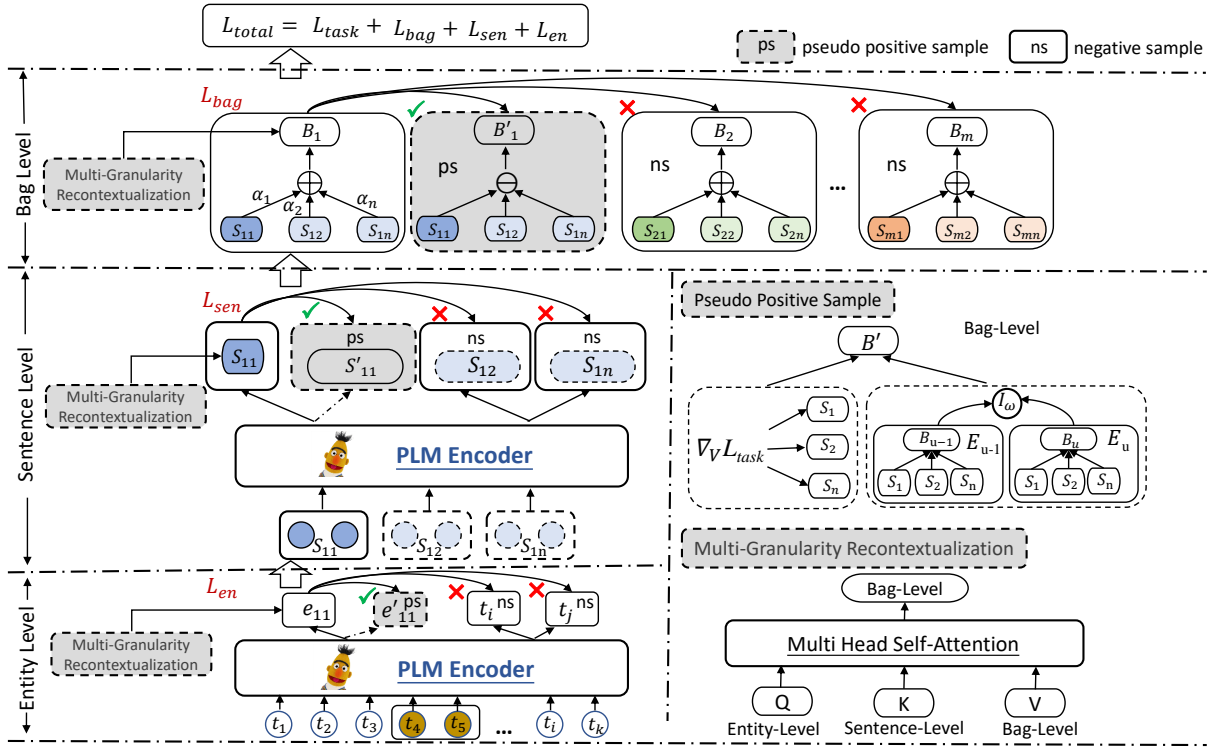
Figure 2: Model overview of HiCLRE. The left part is our model architecture and the right part shows the details on pseudo positive sample construction and multi-granularity recontextualization. (Best viewed in color).

.

embedding processing. ConSERT (Yan et al., 2021) explore four different data augmentation strategies (i.e. adversarial attack, token shuffling, cutoff and dropout) to generate views in BERT (Devlin et al., 2019) embedding layer. SimCSE (Gao et al., 2021) applies twice dropout in the forward process to refine the better sentence representation. (3) Data augmentation by external knowledge. ERICA (Qin et al., 2021) enumerates all the entity pairs in the training samples to link the corresponding relation from the external knowledge graph to obtain sufficient augmented data. The mentioned above methods are generally augmenting from the data aspect, ignoring the influence of the changes inside the model during the training process (Zang et al., 2020; Zou et al., 2020). Hence, we propose a hierarchical contrastive learning model to capture the global structure information and fine-grained interaction within the levels.

3 Methodology

3.1 Model Overview and Notations

The main architecture of our model is shown in Figure 2. The HiCLRE mainly includes two components. (1) Multi-Granularity Recontextualization

aims to integrate the importance of cross levels to determine what valuable representation should be extracted in the target level. (2) Dynamic Gradient Adversarial Perturbation is proposed for specific levels to enhance the internal semantics via constructing the pseudo positive samples.

In HiCLRE, each sentence of input samples is consisted of certain tokens $S_{ij} = (t_{i1}, t_{i2}, \cdots, t_{ik})$, where S_{ij} denotes the i-th sentence of bag B_j. k is the total number of tokens in S_{ij} and j represents the bag's index. e_{i1} and e_{i2} are head and tail entity of sentence S_{ij} respectively. Each bag contains n sentences $B_j = (S_{1j}, S_{2j}, \cdots, S_{nj})$. Our model aims to predict the specific relation r_j of bag B_j from $|r|$ relations. d denotes the hidden state dimension of pre-trained language models (PLMs).

3.2 Hierarchical Learning Modeling

We first introduce our hierarchical learning process including sentence-level and bag-level respectively and then describe the Multi-Granularity Recontextualization and Dynamic Gradient Adversarial Perturbation specifically.

3.2.1 Sentence Representation

To be specific, the input of sentence encoder is the token sequence of sentence S_{ij} and its corresponding head entity e_{i1} and tail entity e_{i2}[2]. The textual encoder sums the token embedding, segment embedding and position embedding for each token to achieve its input embedding, and then computes context-aware hidden representations $H = \{h_{t_{i1}}, h_{t_{i2}}, \cdots, h_{e_{i1}}, \cdots, h_{e_{i2}}, \cdots, h_{t_{ik}}\}$:

$$H = \mathcal{F}\left(\{t_{i1}, t_{i2}, \cdots, t_{ik}\}\right) \quad (1)$$

where $\mathcal{F}$ is the PLMs (e.g. BERT) as our encoder and $H \in \mathbb{R}^{k \times d}$. The sentence's embedding is calculated by the hidden representations of head entity, tail entity and the $[CLS]$ tag, which is in the first position of the input sequence to denote the whole semantic of the sentence.

$$h_{S_{ij}} = \sigma([h_{e_{i1}} \parallel h_{e_{i2}} \parallel h_{[CLS]}] \cdot W_S) + b_S \quad (2)$$

where the $\parallel$ means the concatenation operation, $W_S \in \mathbb{R}^{3d \times d}$ is a weight matrix and b_S is the bias. σ denotes the non-linear function.

3.2.2 Bag Representation

In this section, we use a sentence-level attention-based mechanism (Lin et al., 2016) to yield the aggregated bag representation. Let $h_{B_j} \in \mathbb{R}^d$ denotes the bag representation, and which is computed from the sentence's attention weight α_{ij} and hidden representation $h_{S_{ij}}$.

$$h_{B_j} = \sum_{i=1}^{n} \alpha_{ij} h_{S_{ij}} \quad (3)$$

To avoid naively treating each sentence of bags equally, the selective attention mechanism assigns the importance to reduce the noise instance. Each weight α_{ij} is generated by a query-based function:

$$\alpha_{ij} = \frac{\exp(f_{ij})}{\sum_n \exp(f_{ij})} \quad (4)$$

where f_{ij} measures how well the input sentence S_{ij} and the predicted relation r_j matches.

$$f_{ij} = h_{S_{ij}} \mathbf{A_j} \mathbf{r_j} \quad (5)$$

where $\mathbf{A_j} \in \mathbb{R}^{d \times d}$ is a weighted diagonal matrix, and $\mathbf{r_j} \in \mathbb{R}^d$ is the representation of relation r_j

which is mapped from the relation label. The final relation type of bag B_j is predicted:

$$p(r_j \mid h_{B_j}, \theta) = \frac{\exp(O_r)}{\sum_{p=1}^{|r|} \exp(O_p)} \quad (6)$$

$$O_r = \sigma(W_r \cdot h_{B_j}) + b_r \quad (7)$$

where $W_r \in \mathbb{R}^{|r| \times d}$ is trainable transformation matrix and $b_r \in \mathbb{R}^{|r|}$ is the bias. θ denotes bag encoder's parameters. $O_r \in \mathbb{R}^{|r|}$ represents the final output of our model, which is associated with all relation types. Therefore, the relation classification objective function of DSRE task is denoted as:

$$\mathcal{L}_{task} = -\sum_{j=1}^{|r|} \log p\left(r_j \mid h_{B_j}, \theta\right) \quad (8)$$

3.3 Multi-Granularity Recontextualization

The hierarchical learning process described above neglects the explicit interaction of cross levels to refine the better level's representation. Hence, after updating the hidden representations generated by the PLMs, our HiCLRE model attempts to recontextualize the enhanced representations for each level. This is accomplished using a modified Transformer layer (Vaswani et al., 2017) that substitutes the multi-headed self-attention with multi-headed attention between the target level and the other two levels' representations.

Specifically, the underlying calculation process of multi-head self-attention is defined as:

$$\text{Att.}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (9)$$

where $<Q, K, V>$ means query, key, and value respectively. d_k is the dimension of K. For example, if we focus on the enhanced bag-level representation [3], the h_{B_j} is substituted for the value, whereas the sentence-level $h_{S_{ij}}$ and entity-level h_e mean the key and query respectively[4]:

$$h'_{B_j} = \text{MLP}(\text{Att.}(h_e, h_{S_{ij}}, h_{B_j})) \quad (10)$$

where MLP is the linear multi-layer linear function. The similarity calculation (i.e. query and key) acts as the cross-level information interaction attending to the bag-level representation. After the interaction with multi-headed attention, we run a position-wise MLP similar to the standard transformer layer.

---

[2] Entity's representation is calculated by averaging all tokens hidden states of the entity.

[3] The calculation process of other modules is identical for entity level and sentence level. Hence, we take the bag level as an example in the following paper.

[4] Swapping the meaning of Q and K is also permitted.

Next, we concatenate enhanced target level representation with original hierarchical hidden state to obtain an informative level's representation:

$$h_{B_{att_j}} = \sigma([h_{B_j} \parallel h'_{B_j}] \cdot W_{att}) + b_{att} \quad (11)$$

where $W_{att} \in \mathbb{R}^{2d \times d}$ is a weight matrix and b_{att} is the bias. Finally, we leverage the three-level enhanced representation $h_{e_{att_j}}, h_{S_{att_j}}$ and $h_{B_{att_j}}$ to replace the hierarchical hidden representation in the following calculation process.

3.4 Dynamic Gradient Adversarial Perturbation

In addition to considering the interaction of cross levels, the semantic differences of fine-grained relations within the levels can also help models further enhance the context-aware representations. We construct a pseudo positive sample for contrastive learning (Jaiswal et al., 2020) to push the dissimilar relations away. Since the changes of specific-level gradient (Zhang et al., 2020) and the better context-aware semantic can boost the robustness representations, we devise the gradient perturbation and inertia weight memory mechanisms respectively.

3.4.1 Gradient Perturbation

The continuous gradient perturbations pt_{adv} is calculated from the gradient g of the task loss with the parameter V.

$$g_j = \bigtriangledown_V \mathcal{L}_{task}(h_{B_j}; \theta) \quad (12)$$

where V is the representation of the bag's sentences. We differentiate the entity to generate the gradient perturbation for sentence level and the token for the entity level.

$$pt_{adv_j} = \epsilon \cdot \frac{g_j}{\|g_j\|} \quad (13)$$

where $\|g\|$ is the norm of the gradient from the loss function, ϵ is a hyperparameter to control the disturbing degree.

3.4.2 Inertia Weight Memory

With the training epoch increasing, we use the time-sequential information of different granularities to further improve the robustness of internal semantics. Specifically, we add the inertia weight information (Shi and Eberhart, 1998) on the perturbation term, which takes advantage of the difference of representations between the last and the current

epoch. The inertia weight information is denoted as follows:

$$I_w = \frac{T - u}{T} \text{sim} \left(rep_{(u)}, rep_{(u-1)} \right) \quad (14)$$

where T is the total epoch number of the training process and u is the current epoch index. $rep_{(u)}$ can denote the entity, sentence, or bag representation respectively of the u-th epoch. rep is a embedding matrix saving the semantic memory in the order of element index, updated from the second epoch during the training process. Then, we combine the inertia weight information with gradient perturbation for bag level:

$$pt_{adv_j} = \epsilon \frac{g_j}{\|g_j\|} + \frac{T - u}{T} \text{sim} \left(rep_{(u)}, rep_{(u-1)} \right) \quad (15)$$

We add pt_{adv_j} into the bag embedding, and get pseudo positive sample $h'_{B_j} = h_{B_j} + pt_{adv_j}$. Then we randomly sample a bag in the batch act as the negative sample. The positive and negative samples in InfoNCE loss (van den Oord et al., 2018) are replaced by the dynamic gradient perturbations and random bags respectively:

$$\mathcal{L}_{bag}^{info} = -\log \frac{\exp\left(cos\left(h_{B_j}, h'_{B_j}\right)/\tau\right)}{\sum_{k=1}^{m} \mathbb{1}_{[k \neq j]} \exp\left(cos\left(h_{B_j}, h_{B_{k_j}}\right)/\tau\right)} \quad (16)$$

where $\mathbb{1}_{[k \neq j]}$ is an indicator function, τ is a hyper-parameter and cos is the cosine function. Due to the different granularities in the hierarchical framework, we devise different memories for entity-level, sentence-level, and bag-level, respectively.

3.5 Training Objective

In HiCLRE, our training objective contains two components, including the DSRE task loss and the contrastive learning loss. The total loss of contrastive learning is the sum of three-level infoNCE loss. Therefore, the overall objective function is formulated as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{en}^{info} + \lambda_2 \mathcal{L}_{sen}^{info} + \lambda_3 \mathcal{L}_{bag}^{info} + \lambda_4 \mathcal{L}_{task} \quad (17)$$

where λ_l is hyper-parameter and $\sum_{l=1}^{4} \lambda_l = 1$, denoting the weight of each components.

4 Experiments

4.1 Datasets and Baselines

We evaluate our HiCLRE model on three DSRE datasets, including NYT10 (Riedel et al., 2010),

Models	NYT10					GDS				
	AUC	P@100	P@200	P@300	P@M	AUC	P@500	P@1000	P@300	P@M
Mintz	10.7	52.3	50.2	45.0	49.2	-	-	-	-	-
PCNN-ATT	34.1	73.0	68.0	67.3	69.4	79.9	90.6	87.6	75.2	84.5
MTB-MIL	40.8	76.2	71.1	69.4	72.2	88.5	94.8	92.2	87.0	91.3
RESIDE	41.5	<u>81.8</u>	75.4	<u>74.3</u>	<u>77.2</u>	89.1	94.8	91.1	82.7	89.5
REDSandT	42.4	78.8	75.0	73.0	75.3	86.1	95.6	92.6	84.6	91.0
DISTRE	42.2	68.0	67.0	65.3	66.8	89.9	97.0	93.8	87.6	92.8
CIL	<u>43.1</u>	81.5	<u>75.5</u>	72.1	76.9	<u>90.8</u>	<u>97.1</u>	<u>94.0</u>	<u>87.8</u>	<u>93.0</u>
HiCLRE(ours)	**45.3**	**82.0**	**78.5**	74.0	**78.2**	**95.5**	**99.6**	**98.4**	**98.3**	**98.8**

Table 1: General experimental results of HiCLRE and baselines on NYT10 and GDS datasets.

GDS (Jat et al., 2017), and KBP (Ling and Weld, 2012). Table 4 shows the detailed statistics. NYT10 is annotated from the New York Times and aligned to Freebase and NYT10-M removes the noisy relation types manually from NYT10. GDS is extracted from human-judged Google Relation Extraction corpus. KBP is constructed over the newswire and web text from the corpus, which is used in the yearly TAC Knowledge Base Population challenges (Ji et al., 2010). Statistics of four datasets are showed in Appendix A.

Mintz (Mintz et al., 2009) concatenates various features of sentences to train a multi-class logistic regression classifier. PCNN-ATT (Lin et al., 2016) proposes a selective attention-based piecewise CNN to get sentence embeddings. MTB-MIL (Soares et al., 2019) proposes a Matching the Blanks method to learn the sentences' representation by the entity linked text. RESIDE (Vashishth et al., 2018) exploits the information of entity type and relation alias to add a soft limitation for relation classification. REDSandT (Christou and Tsoumakas, 2021) employs the PLMs to focus on instance embedding, aggregating the representations to the attention modules. DISTRE (Alt et al., 2019) combines the selective attention to its Transformer-based model. CIL (Chen et al., 2021a) proposes a contrastive instance learning method under the MIL framework.

4.2 Evaluation Metrics

Following the previous works (Chen et al., 2021b), we adopt the five general evaluation metrics in DSRE task to evaluate the performance, including **AUC**, **P@N** and **P@M**. Specifically, AUC (i.e. Area Under Curve) depicts the area under the ROC curve [5]. P@N refers to the P@100, P@200 and

P@300 used in the metrics, denoting the top 100, top 200 and top 300 precision respectively. P@M is the mean value of the above three P@N results.

4.3 Parameter Settings

The underlying encoders of the entity level and sentence level are implemented by BERT_base (Devlin et al., 2019). The backbone encoder contains 12 Transformer layers and 12 self-attention heads, generating 768 hidden units for each token context-aware representation. During the training stage, we set the model's learning rate as {1e-5, 2e-5, 2e-7}. We choose AdamW (Loshchilov and Hutter, 2017) as our model's loss optimizer, which weight decay is 1e-5 and learning rate is 0.1. The max epoch is set to 5. We find the best hyper-parameter of temperature τ is 0.05, the λ set is {0.4, 0.4, 0.1, 0.1} and ϵ is 2. We show the important hyper parameters' searching results at Appendix C.

4.4 General Experimental Results

We first evaluate our HiCLRE model in the NYT10 and GDS that are popular used datasets in the DSRE task. Table 1 shows the overall performance on the NYT10 and GDS datasets. From the results[6], we can observe that (1) On both two datasets, the performance of our HiCLRE model outperforms all the strong baseline models significantly on the four metrics, achieving a new state-of-the-art result. (2) The performance of HiCLRE is greatly improved compared with the strongest baseline in two distantly supervised datasets (i.e. +2.2 AUC / +4.7 AUC). Meanwhile, we find the results of other four metrics are also increasing consistently. In general, it can be seen from Table 1 that the multi-granularity recontextualization for cross levels interaction and the dynamic gradient-based ad-

---

[5]ROC curve is plotted by false positive rate and true positive rate.

[6]CIL is the SOTA model in these datasets, whereas the source code is not provided so far. Hence, we reproduce the CIL model and report the performance in the test set.

versarial perturbation for specific levels can improve the performance greatly. Some general cases also prove the effectiveness in Appendix B.

The baselines and HiCLRE's overall PR-cureve is illustrated in Figure 3. From the curve, we can observe that (1) Our HiCLRE shows higher precision and recall results compared to other strong baselines. (2) Although the curve initially fluctuates quite a bit, both metrics of HiCLRE are basically stabilized at a relatively large gap during the training process. We conjecture that the difference of hierarchical context-aware representation is not obvious at the beginning of the model's training and the stored representations for inertia weight memory of specific levels do not exist in the first training epoch. During the training process, the mentioned above two learning problems tend to be stable and the performance of these two metrics is continuously performing better.
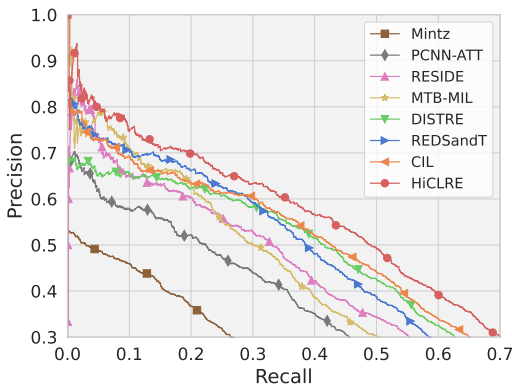

Figure 3: PR-curves of HiCLRE and other baselines on NYT10 dataset. (Best viewed in color)

Models	NYT10-M			KBP		
	AUC	F1	P@M	AUC	F1	P@M
PCNN-A	41.9	32.0	68.6	15.4	31.5	32.8
DISTRE	35.7	31.4	65.1	22.1	37.5	46.4
CIL	56.0	34.3	75.9	29.5	41.6	47.3
HiCLRE	**61.4**	**36.9**	**88.0**	**46.1**	**61.0**	**56.4**

Table 2: Experiment results on human-annotated datasets.

4.5 Evaluation on Human Annotated Dataset

Due to the inevitable annotated errors of distantly supervision assumption, we further evaluate our model on the human-annotated high-quality relation extraction datasets, including NYT10-M and KBP. The performances of baselines and HiCLRE are shown in Table 2. The result shows that Hi-

CLRE can significantly outperform the three strong baselines especially the AUC metric reaches 46.1, which improves about 50% ($29.5 \mapsto 46.1$) performance than CIL (Chen et al., 2021a) on the KBP dataset. This phenomenon implies that our model possesses a steady generalization ability to other analogous relation extraction datasets.

5 Detailed Analysis of HiCLRE

5.1 Ablation Study

To verify the effectiveness of various modules in our HiCLRE model, we conduct ablation study experiments on the NYT10 dataset. Specifically, we remove the following argued contributions in turn to evaluate the performance, including multi-granularity recontextualization, three-level contrastive learning loss, and the data augmentation strategies in each level. The final results are shown in Table 3. From the results, we conclude that (1) The context-aware representation interactions for cross levels and the enhanced internal semantics representations for specific level are essential, dropping -1.8% and -2.7% point on the AUC metric respectively. (2) We also find the sentence-level data augmentation skills for our HiCLRE model are the most important (e.g. -4.9% and -2.6% on AUC) compared to the other two levels. The possible reason may be that the sentence granularity is the fundamental input granularity for the DSRE task including the term "bag" is also constructed by choosing the sentences with identical entity pairs.

Methods		AUC	F1	P@M
HiCLRE		45.3	49.5	78.2
-Multi-Gra. Recon.		43.5	47.1	76.4
-Three-level CL Loss		42.6	47.9	70.8
-Bag Level	-bag gradient	43.9	48.1	75.4
	-bag memory	42.4	48.6	72.3
-Sentence Level	-sen. gradient	40.4	48.9	73.9
	-sen. memory	42.7	48.1	67.2
-Entity Level	-en. gradient	43.0	48.4	70.1
	-en. memory	43.1	46.7	73.2

Table 3: Ablation study of HiCLRE on NYT10. "-" means removing the module behind.

5.2 The Influence of Multi-Granularity Recontextualization

Figure 4 shows the comparison of final stable results and speed of convergence between the

multi-granularity recontextualization and single-granularity[7] on the NYT10 dataset. we can observe that (1) multi-granularity recontextualization converges faster to not only stable but also better results. (2) When the final performance of the model converge stably, our multi-granularity recontextualization have less jitter amplitude making our model more robust.
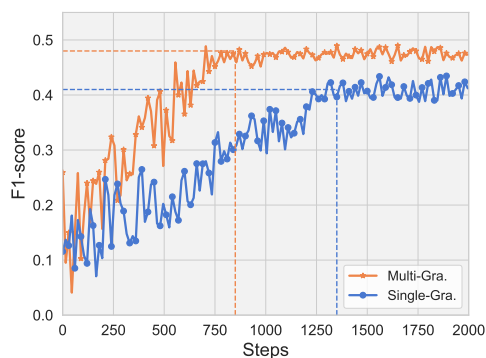
bag relation: /business/person/company
S_3: In 1998, Steve Jobs asked Cook to join Apple.
S_4: Under Cook's leadership, Apple has increased its donations to charity.
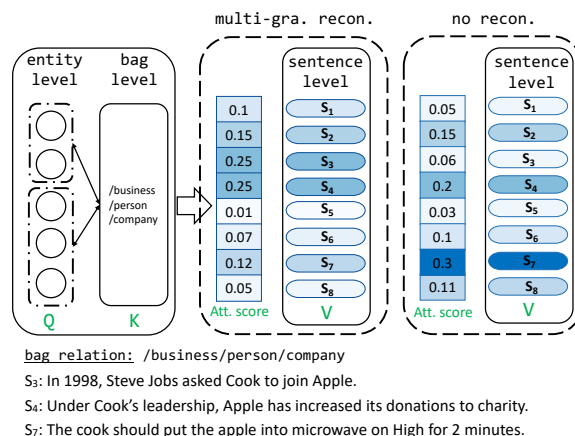S_7: The cook should put the apple into microwave on High for 2 minutes.

Figure 5: The heat maps of our multi-headed attention mechanism (ours) and no attention mechanism (right) among cross levels. (Best viewed in color)

Figure 4: The comparison including convergence speed (i.e. steps) and performance (i.e. F1-score) of training process between our multi-granularity cross-level attention (Multi-Gra.) and single-granularity (Single-Gra.). (Best viewed in color)

Figure 5 shows the two heat maps of the module with and without attention calculations, proving our multi-granularity recontextualization mechanism is effective for denoising the redundant sentences. For example, we take the sentence-level representations act as "V" value. Our multi-granularity recontextualization mechanism can achieve the higher attention scores (e.g. S_3 and S_4) for the important sentences in a bag, whereas the no recontextualization models incorrectly assign the highest attention score (e.g. S_7) to the noisy sentences. This phenomenon indicates that this mechanism has a better ability to filter noisy sentences.

5.3 The Influence of Gradient-based Data Augmentation

To further prove our data augmentation skill of contrastive learning is effective, we choose the other three strategies (i.e. randomly deleting a word, twice dropout, and randomly noise) to perform the evolution process of representation learning space.

Following the previous works (Wang and Isola, 2020), we treat the pseudo sample as the positive

instance and a randomly chosen instance from the batch as the negative instance to calculate alignment and uniformity. Then, we plot the transformation of the align-uniform points in Figure 6. The lower alignment and uniformity results indicate the better context-aware representations of the contrastive learning process. Compared to the other positive sample generation skills, our dynamic gradient adversarial perturbation module reduces the alignment and uniformity metrics steadily to the lower value and faster speed.
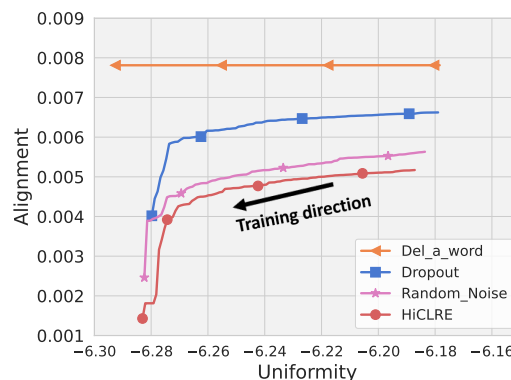
Figure 6: Results comparison of HiCLRE and other data augmentation skills in terms of alignment and uniformity. The arrows indicate the training direction. (Best viewed in color).

.

6 Conclusion

In this paper, we propose HiCLRE, a hierarchical contrastive learning framework for distantly supervised relation extraction. Multi-Granularity Recontextualization module of HiCLRE utilizes a multi-head self-attention mechanism to transmit

---

[7]The single-granularity means just facility original representation of each level without combining the multi-headed attention mechanism.

the information across three levels. Dynamic Gradient Adversarial Perturbation module combines the gradient perturbation with inertia memory information to construct better pseudo positive samples for contrastive learning. Experiments show the effectiveness of HiCLRE against the strong baseline models in various DSRE datasets.

Acknowledgements

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in relation extraction. In *ACL*, pages 1534–1545.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *ACL*, pages 1388–1398.

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019. Combining distant and direct supervision for neural relation extraction. In *NAACL*, pages 1858–1867.

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021a. CIL: contrastive instance learning framework for distantly supervised relation extraction. In *ACL*, pages 6191–6200.

Tao Chen, Haochen Shi, Liyuan Liu, Siliang Tang, Jian Shao, Zhigang Chen, and Yueting Zhuang. 2021b. Empower distantly supervised relation extraction with collaborative adversarial training. In *AAAI*, pages 12675–12682.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2021. Distantly supervised relation extraction with sentence reconstruction and knowledge base priors. In *NAACL*, pages 11–26.

Despina Christou and Grigorios Tsoumakas. 2021. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, pages 62574–62582.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, pages 297–304.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, pages 541–550.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *CoRR*, abs/2011.00362.

Sharmistha Jat, Siddhesh Khandelwal, and Partha P. Talukdar. 2017. Improving distantly supervised relation extraction using word and entity based attention. In *NIPS*.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *TAC*, volume 3, pages 3–3.

Khalid Al Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *AAAI*, pages 7367–7374.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Hai Zhao. 2019. GAN driven semi-distant supervision for relation extraction. In *NAACL*, pages 3026–3035.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *ACL*.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *AAAI*.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. Rikinet: Reading wikipedia pages for natural question answering. In *ACL*, pages 6762–6771.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. SENT: sentence-level distant relation extraction via negative training. In *ACL*, pages 6201–6213.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*, pages 1003–1011.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. DSGAN: generative adversarial training for distant supervision relation extraction. In *ACL*, pages 496–505.

Yujia Qin, Yankai Lin, Ruyuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. In *ACL*, pages 3350–3363.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML*, pages 148–163.

Apoorv Saxena, Aditay Tripathi, and Partha P. Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *ACL*, pages 4498–4507.

Yuhui Shi and Russell C. Eberhart. 1998. Parameter selection in particle swarm optimization. In *Evolutionary Programming*, pages 591–600.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *ACL*, pages 2895–2905.

Sen Su, Ningning Jia, Xiang Cheng, Shuguang Zhu, and Ruiping Li. 2018. Exploring encoder-decoder model for distant supervised relation extraction. In *IJCAI*, pages 4389–4395.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP*, pages 455–465.

Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *ACL*, pages 2713–2722.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. pages 2579–2605.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha P. Talukdar. 2018. RESIDE: improving distantly-supervised neural relation extraction using side information. In *EMNLP*, pages 1257–1266.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Exploiting the syntax-model consistency for neural relation extraction. In *ACL*, pages 8021–8032.

Chao Wang and Hui Jiang. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *ACL*, pages 2263–2272.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP*, pages 6381–6387.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *ACL*, pages 1476–1488.

Shanchan Wu, Kai Fan, and Qiong Zhang. 2019. Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector. In *AAAI*, pages 7273–7280.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *ACL*, pages 5065–5075.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *EMNLP*, pages 1456–1466.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *NAACL*, pages 2810–2819.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *ACL*, pages 6066–6080.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

Wei Emma Zhang, Quan Z. Sheng, Ahoud Abdulrahmn F. Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM*, pages 24:1–24:41.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2020. A reinforced generation of adversarial examples for neural machine translation. In *ACL*, pages 3486–3497.

A   Datasets Statistics

Dataset	# Rel.	# Train	# Test	Test Type
NYT10	58	522,611	172,448	DS
GDS	5	18,328	5,663	Partly MA
NYT10-M	25	417,893	11,085	MA
KBP	12	87,940	288	MA

Table 4: Statistics of four datasets. Rel.: relation, DS: distantly supervised and MA: manually annotated.

B   Case Study

We enumerate several representative examples in Figure 7 to further explore why our model can work in the distantly supervised scenario. In the left part of the figure, there are two bags containing the different entity pairs (i.e. ⟨"Bill Gates", "Microsoft" ⟩ and ⟨"Robert Walter", "Cardinal Health" ⟩). Previous works ignore the consideration of the representation interaction in a specific levels and cross levels, which may be hard to predict similar or difficult instances. For example, sentences of bag B_1 are always classified into the relation "*major_shareholders_of*". Although these two relations (i.e. "*major_shareholders_of*" and "*/person/company*") are pretty similar to each other, none of the four sentences' semantics in B_1 represent the meaning of "*major_shareholders_of*". In particular, after the context-aware representations interaction via cross levels and specific levels, HiCLRE can correctly predict the bag to the ground-truth label; likewise, the bag B_2 is in the same situation.

In the right part of the figure, we demonstrate an example of that HiCLRE can pull the correlated instance closely and push the uncorrelated instance away. We reduce the dimension of bag examples' representations by t-SNE (van der Maaten and Hinton, 2008) and show the example results in the coordinate system. R^* is the target instance to be classified, the symbol "$+$" represents the degree of relevance and "$-$" represents the degree of irrelevance. HiCLRE can pull the related sample R^+ to R^{+++} (i.e. closer to R^*), while pushing the uncorrelated sample R^- to R^{--} (farther from R^*). This phenomenon is own to the design of gradient-based perturbation, which gives significant enhancement to interactions in a specific level.

C   The influence of important hyper parameters

We experiment with our model on the NYT10 dataset with four important hyper-parameters, and discover a suitable parameters' combination to reach better performance.

| Bag	Sentence	Wrong Prediction	HiCLRE(ours) Prediction
B_1	• If **Bill Gates** had to worry about health insurance would he have started **Microsoft** ? • **Bill Gates** will be involved , Mr. Trump said , with a new **Microsoft** product . • **Microsoft** will invest $ 1.7 billion in India over the next four years, its chairman , **Bill Gates** , said Wednesday . • Three decades after he started **Microsoft** with the dream of placing a personal computer in every home and business , **Bill Gates** said that he would leave his day-to-day role there in two years .	/business /company_ shareholder /major_shar eholders_of	/business /person /company
B_2	• **Robert Walter** retired from **Cardinal Health** in June 2008. • **Robert Walter** is setting down his memories of **Cardinal Health**'s life. • **Cardinal Health** reported $151000 in personal aircraft use last year for **Robert Walter**.	/business /company /founders	/business /person /company
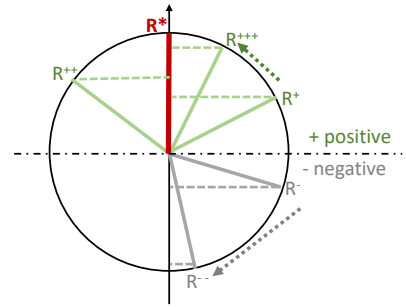
Figure 7: Examples of cases in our experiments. The left table means the comparison of predicted labels. The right figure shows the Dynamic Gradient Adversarial Perturbation module's working process.(Best viewed in color).
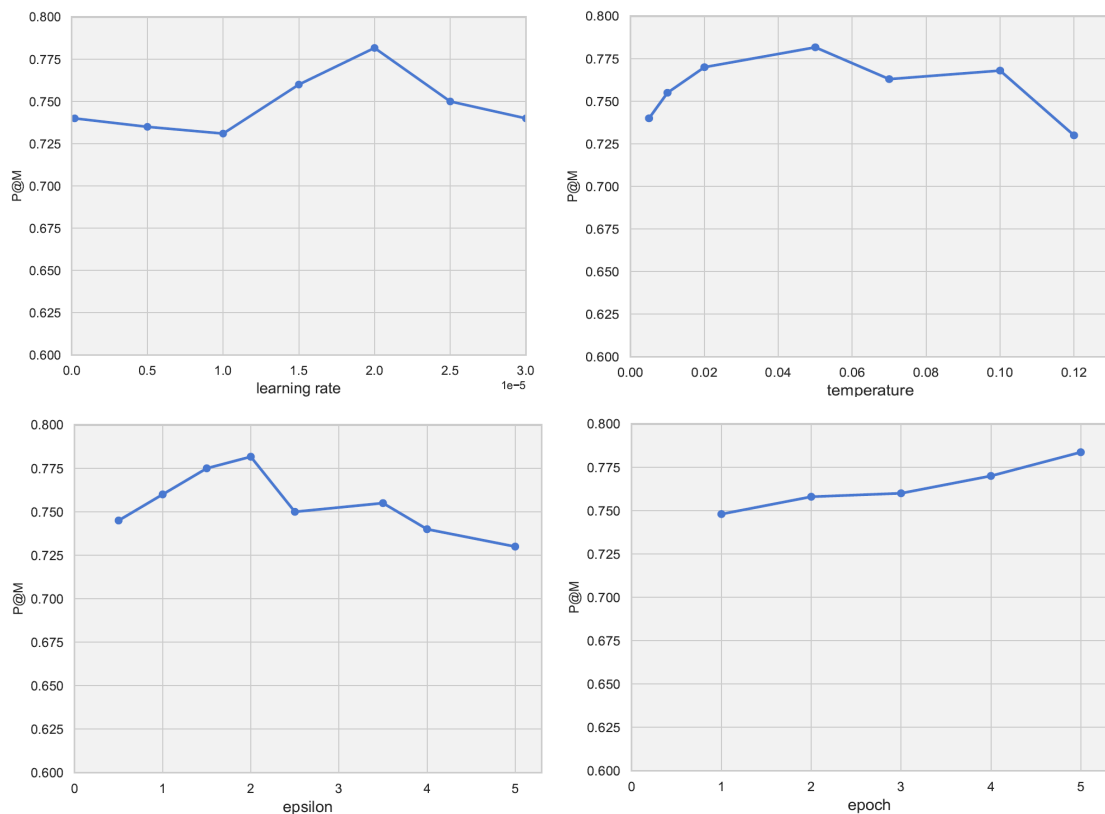
.

Figure 8: The influence of four important hyper-parameters on the NYT10 dataset. (Best viewed in color).

.