NS-Hunter: BERT-Cloze Based Semantic Denoising for Distantly Supervised Relation Classification

Tielin Shen, Daling Wang, Shi Feng, Yifei Zhang
School of Computer Science and Engineering, Northeastern University, Shenyang, China
{shentielin, wangdaling, fengshi, zhangyifei}@cse.neu.edu.cn

Abstract

Distant supervision can generate large-scale relation classification data quickly and economically. However, a great number of noise sentences are introduced which can not express their labeled relations. By means of pre-trained language model BERT's powerful function, in this paper, we propose a BERT-based semantic denoising approach for distantly supervised relation classification. In detail, we define an entity pair as a source entity and a target entity. For the specific sentences whose target entities in BERT-vocabulary (one-token word), we present the differences of dependency between two entities for noise and non-noise sentences. For general sentences whose target entity is multi-token word, we further present the differences of last hidden states of [MASK]-entity (MASK-lhs for short) in BERT for noise and non-noise sentences. We regard the dependency and MASK-lhs in BERT as two semantic features of sentences. With BERT, we capture the dependency feature to discriminate specific sentences first, then capture the MASK-lhs feature to denoise distant supervision datasets. We propose NS-Hunter, a novel denoising model which leverages BERT-cloze ability to capture the two semantic features and integrates above functions. According to the experiment on NYT data, our NS-Hunter model achieves the best results in distant supervision denoising and sentence-level relation classification.

Keywords: Distant supervision, relation classification, semantic denoising

1 Introduction

Relation classification (RC) is a fundamental task in natural language processing. The goal of RC (Zelenko et al., 2003) is to identify the relation type in a sentence for a given entity pair, which is particularly important for the construction of knowledge bases. In recent years, deep learning has performed very well in relation extraction, but the technique needs a great number of labeled data, which is very expensive for manual tagging. In order to obtain a large amount of labeled RC data quickly and cheaply, distant supervision (DS) (Mintz et al., 2009) was proposed to automatically generate data by aligning a knowledge base with an unlabeled corpus. It is built on a weak assumption that if an entity pair has a relation in a knowledge base, all the sentences containing this entity pair will express the corresponding relation and exist in the dataset as a bag (Mintz et al., 2009). Based on such an assumption, a large number of noise sentences are generated by DS because many sentences can not express their labeled relation in fact. For example, in Figure 1, sentence S-2 can not express the labeled relation "founder". These noise sentences such as S-2 will cause error propagation and may significantly reduce the performance of RC model.

Since the pre-trained language model BERT was put forward (Devlin et al., 2018), it has performed very well in the fully supervised RC datasets such as SemEval 2010 task 8 (Soares et al., 2019; Wu and He, 2019). Different from DS, it is manually labeled, so there is no noise-sentences in it. Following these two works, even if we only keep two entities and delete the other parts of the sentence in test set, we can still get the F1-value of 49.99% (89.2% in MTB (Soares et al., 2019)) in 19-class fully supervised RC task. Only keeping entity pair in every sentence of test set can get such a high F1-value, which shows that BERT-based RC model will pay more attention to the entity pair itself rather than other words of

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, August 13 - 15, 2021.

Entity Pair and Labeled Relation

Entity1	Entity2	Relation
Jobs	Apple	founder

Two Example Sentences

S-1	Jobs was the co-founder and CEO of Apple.
S-2	During Jobs' tenure, Apple released four iPhones.

Figure 1: Two sentences generated by DS method.

the sentence. So, the model can not effectively discriminate noise and non-noise sentences, because in a bag generated by DS-method, noise and non-noise sentences have the same entity pair. Therefore, this method can not be directly used in sentence-level DS-RC.

Predicting masked word is one of the two pre-training tasks of BERT. Cui et al. (Cui et al., 2020) showed that there is a great deal of commonsense knowledge in BERT. After our verification, BERT can predict most blanks in general texts, which enables us to identify the noise-sentences in DS dataset.

In this paper, we propose NS-Hunter, a novel denoising model for DS-RC, which leverages BERT-cloze to capture semantic features of noise sentences in DS dataset (Short for **N**oise **S**entence **Hunter**). We define the entity pair in each sentence as a **source entity** and a **target entity**, and the rest of the sentence as **relation pattern**. Here the source entity is known, but the target entity needs to be predicted which may be head entity or tail entity. Our NS-Hunter is based on the following assumption: **in a non-noise sentence, the correct prediction of the target entity requires the both attendance of source entity and relation pattern.** The assumption restricts the dependency of the source entity and target entity based on the relation pattern marked, which means for a sentence, if the target entity can be predicted only based on one of the either source entity or relation pattern, it will be regarded as a noise sentence. We regarded the dependency of the source entity and the target entity as the first semantic feature for detecting noise sentence in DS datasets.

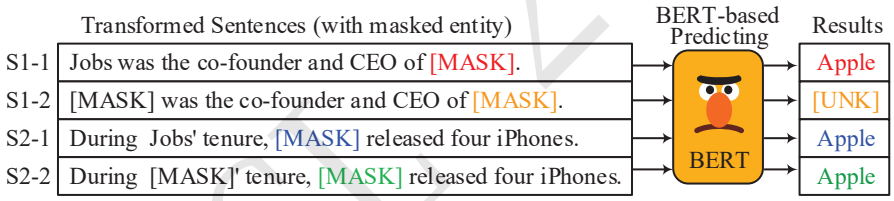

Figure 2: An example of identifying noisy sentences.

According to the function of BERT, if a word in a sentence was replaced by [MASK] and then the revised sentence was input to BERT, BERT can make a reasonable prediction for the word at the position of [MASK] according to other words in the sentence. To explain the rationality of our assumption, we transformed the S-1 in Figure 1 into S1-1 and S1-2 in Figure 2, and also the S-2 in Figure 1 into S2-1 and S2-2 in Figure 2. For S1 in Figure 1, when we do not mask Jobs, i.e. S1-1 in Figure 2, BERT can predict Apple. When we mask Jobs with [MASK], i.e. S1-2 in Figure 2, BERT can not predict Apple. That means Jobs and Apple have closely related dependency in S-1, so we think S-1 can express the labeled relation. We use the same method to judge S-2, whether we mask Jobs (i.e. S2-2) or not (i.e. S2-1), BERT can predict Apple, which shows that in S-2, Jobs and Apple are loosely related, so we think S-2 can not express the labeled relation, and it is a noise sentence. It is based on the first semantic feature, i.e. dependency feature, we can recognize a part of noise-sentences.

However, this method is only applicable to the sentences whose target entities in BERT-vocabulary (one-token word). In order to make our model can deal with multi-token word, we build some noise reducers, which are actually binary classifiers based on BERT. Their training sets and development sets are from the part that can be discriminated in the original training sets with our dependency feature, while their test sets are the whole original training set, so we can denoise the original training set for RC. For S2-1 and S2-2 in Figure 2, the prediction results of [MASK] are all Apple, but the prediction

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109–1120, Hohhot, China, August 13 – 15, 2021.

results of [MASK] in S1-1 and S1-2 are different. So, we consider that there are semantic difference in the [MASK]-represention between noise and non-noise sentences, which can be captured by fully-connected layer. For the reason, we do not use the commonly used [CLS] feature, but concatenate last hidden states of [MASK] in transformed sentences as the feature (we call it as MASK-lhs), and regard MASK-lhs as the second semantic feature. We utilize MASK-lhs feature to denoise general sentences.

To the best of our knowledge, this is the first work of presenting semantic feature differences between noise and non-noise sentences, and implementing semantic denoising based on the features. We use the sentence-level test set in ARNOR tagged by Jia et al. (Jia et al., 2019) from NYT dataset for sentence-level evaluation of our NS-Hunter.

Overall, our contributions can be summarized as follows:

- We propose a novel model NS-Hunter, which denoises the datasets in DS for RC by leveraging BERT-cloze ability to capture our proposed two noise semantic features (dependency feature and MASK-lhs feature).
- The NS-Hunter we proposed is independent of RC, and it is a plug and play denoiseing model, which can be applied to any existing RC model. We verify the denoising ability of NS-Hunter on CNN (Zeng et al., 2014) and BERT (Devlin et al., 2018).
- We conduct experiments by using ARNOR dataset from NYT. The results show our NS-Hunter model achieves state-of-the-art results.

2 Related Work

Neural network based models have performed very well in RC (Wang et al., 2016). However, training effective neural classifiers requires a large amount of labeled data, which is usually hard to obtain. DS (Mintz et al., 2009) provides a way to create massive weakly labeled data for RC.

Many studies train RC model in DS by applying multi-instance learning (MIL) to reduce the impact of noise-sentences (Lin et al., 2016; Liu et al., 2017; Ji et al., 2017), which relaxes the label of each instance to a bag of sentences containing the same entity pair. Some MIL-based studies introduce adversarial training (Han et al., 2018). MIL assumes at least one sentence in a bag was labeled correctly. When all sentences in a bag are noise-sentences, MIL still suffers from noise (Qin et al., 2018b; Li et al., 2019b). Moreover, these MIL-based approaches are designed and tested for a pair of entities (Li et al., 2019a; Qu et al., 2019), and they are not suitable for sentence-level RC. Alternatively, some studies evaluate and select training instances individually without relying on the at-least-one assumption (Feng et al., 2018; Jia et al., 2019; Qin et al., 2018a; Zeng et al., 2018). They usually rely on the classification effect to denoise. It is difficult to measure the denoising ability alone, and some of them need the pure data labeled manually (Pershina et al., 2014; Beltagy et al., 2019).

Our approach uses the commonsense knowledge in the pre-training language model (PLM) to measure whether the two entity pairs are closely related, does not rely on assumption of MIL and needs no manually labeled data. Our denoising model is independent of classification process. It is plug and play, so can be used with all the above models.

Recently, more and more PLMs have adopted the method of predicting masked word to learn grammar and semantics, such as Roberta (Liu et al., 2019b), Electra (Clark et al., 2020), ERNIE (Zhang et al., 2019) etc. With more and more parameters and larger corpus for training, it can be expected that their cloze-accuracy will be higher and higher and the DS-RC model based on these PLMs will have stronger denoising ability.

Moreover, CASREL (Wei et al., 2019) is BERT-based RC model, they extract entities and relations jointly. This model use DS datasets, but the sentences labeled as NA were deleted. So, we can not compare the denoising ability of NS-Hunter with it.

3 Model

In this paper, we propose NS-Hunter, a novel BERT-based denoising model for DS-RC shown as Figure 3. For a DS dataset with n classes and a NA (no relation) class, we construct a noise reducer with

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, Augest 13 - 15, 2021.

BERT for every class except the NA. Finally, the denoised training set is composed of n classes after denoising and a NA class in the original training set. In Figure 3, the left is the overall framework and the right is the detail of training k-th noise reducer for Class k ($1 \leq k \leq n$) dataset.

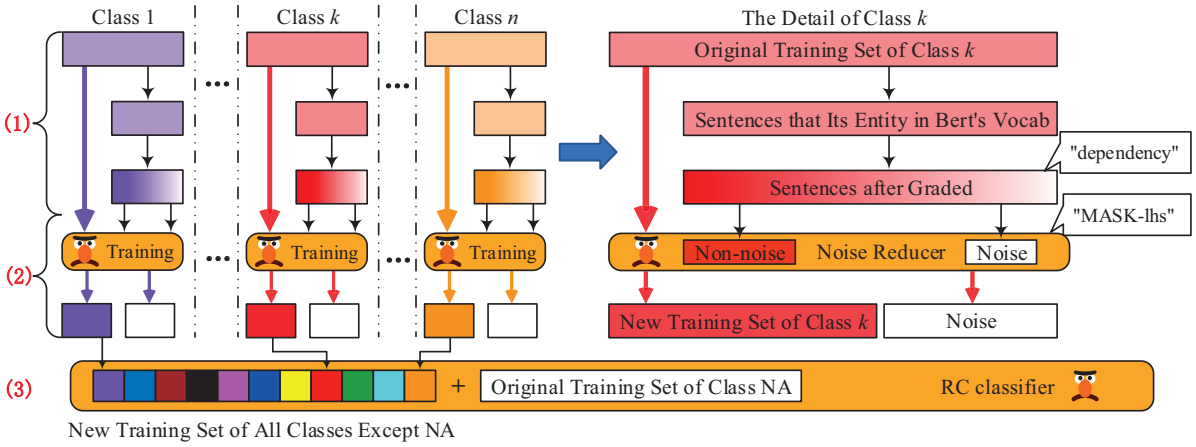

Figure 3: NS-Hunter model. For each class, the lighter color on top means dataset that mixed with noise sentences and non-noise sentences, the gradual color in the middle means the sorted sentences, and the dark color below means the non-noise sentences we extracted. We do the same operation for each class and take the non-noise sentences of each class and the sentences labeled NA in the original dataset as the new training set.

NS-Hunter consists of three parts (see (1)∼(3) on the left side of Figure 3). The first part is to discriminate the sentences whose target entities in BERT-vocabulary, then we get n small-scale pure binary datasets, which is labeled as noise and non-noise. The second part is to train noise reducers with n small-scale datasets. In this part, we design a novel feature MASK-lhs that can effectively capture the semantic differences between noise and non-noise sentences. It should be noted that the first two parts are for the single relation class, that is why we train a noise reducer for each class. The third part is the application of the denoised training set. After denoising, we get a training set with the same form as the original training set, but more pure. We introduce each part in Section 3.1∼Section 3.3, and finally give the algorithm description of NS-hunter.

3.1 Source Entity and Target Entity

The source entity and the target entity are mentioned above, and here we need the target entity existing in BERT-vocabulary when discriminating noise sentences. We assume that in a non-noise sentence, the correct prediction of the target entity requires the both attendance of source entity and relation pattern. Therefore, the target entity must be semantically predictable compared to the source entity. For the example entity pair "Europe" and "Norway" labeled "contains", there is a sentence "Norway is a country in northern Europe" including this entity pair. The relation pattern of this sentence is "* is a country in northern *". However, from the perspective of semantics, it is very difficult to predict "Norway" based only on the relation pattern and "Europe". Obviously, the relation of "Norway" and "Europe" is 1-to-many, here "many" means "Europe" contains many countries, such as "Finland" and "Sweden", which are also semantically reasonable. If "Norway" is the target entity, our NS-Hunter will judge this sentence as noise sentence, but this sentence can actually express "contains" relation. Therefore, "Europe" is the correct target entity, that is to say, we will select a entitiy with a larger scope in the 1-to-many relation as the target entity. According to this method, we can label a source entity and a target entity for each class of dataset. In addition, if the relation between the two entities is 1-to-1 or many-to-many, we can select any one as the target entity.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, August 13 - 15, 2021.

3.2 Discrimination on Dependency Features

There are many entity-pairs with at least two relations (EPO) in the dataset. For example, Biden and the United States have both "place of birth" and "president" relation. According to the DS method, the sentence "Biden is the president of the United States" is labeled as the relation "place of birth" in the dataset. If we only build a noise reducer for the training set, this sentence will be considered as non-noise sentence because this sentence can express the relation of "president". However, according to its "place of birth" label, it is a noise sentence because it can not express its labeled relation "place of birth". As shown in Figure 3, we need to build a noise reducer for each class to avoid the influence of EPO on denoising. According to DS method, there is no noise sentence in the NA class, so we do not denoise the NA.

According to dependency features, in a non-noise sentence, the correct prediction of the target entity requires the attendance of both source entity and relation pattern, so we believe the three parts (source entity, target entity and relation pattern) of this sentence are closely related. If the target entity can be predicted only based on either relation pattern (mask source entity with [UNK]) or source entity (delete relation pattern), we think that the three parts of this sentence are loosely related and this sentence is noisy. When predicting masked word in the pre-training of BERT, the last hidden states of [MASK] will pass through a fully connected layer shaped (768, 30522), the numbers in the output represent the possibility that each word in the BERT-vocabulary may appear in the [MASK] position. Therefore, when we predict target entity, we take the corresponding number of the target entity as the prediction score shown as Formula 1. That is why we can only discriminate the sentence whose target entity is one-token word at the beginning.

$$G = g\left(s, en_t\right) \tag{1}$$

where *s* is transformed sentence, en_t is target entity, *g* denotes function based on BERT and *G* is corresponding number of the target entity.

As mentioned above, we can discriminate a part of original training set and train the noise reducers with the pure datasets after discriminating. Here we believe that if the higher *G* is when source entity and relation pattern attend together, and the lower *G* is when source entity or relation pattern attend alone, the higher the possibility that the sentence is not a noise sentence. We use Formula 2 to quantify this possibility.

$$G_s = g\left(en_s + rp, en_t\right) - g\left(rp, en_t\right) - f \tag{2}$$

where G_s is the possibility that a sentence is not a noise sentence, en_s is source entity, en_s+*rp* is transformed sentence such as S1-1 in Figure 2 (target Apple) and means the attendance of both source entity and relation pattern. Moreover, *rp* represents the attendance of only relation pattern such as S1-2 in Figure 2 (target Apple), and *f* represents the probability of predicting the target entity based only on the source entity's attendance.

$$f = max(g\left(en_s^{mp}, en_t\right), g\left(en_s^{mr}, en_t\right)) \tag{3}$$

where en_s^{mp} and en_s^{mr} are artificial sentences, en_s^{mp} is "en_s [MASK]" and en_s^{mr} is "[MASK] en_s". In Figure 1, if we target Apple, en_s^{mp} is "Jobs [MASK]" and en_s^{mr} is "[MASK] Jobs".

In this way, we can grade and sort the sentences whose target entities in BERT-vocabulary. In order to improve the confidence of binary datasets, we discard the middle parts of the sentence sets and take the first *n* and last *n* of the sentence sets as positive and negative samples to train the noise reducer of each class.

3.3 Denoising and Classification

After discriminating, in each class of the training set, we get a binary dataset including noise sentences and non-noise sentences. These datasets only contain one-token entities, so we trained a noise reducer for each class to discriminate noise sentences with multi-token entities. We stated that BERT can't be directly used for sentence level denoising in Section 1, so we designed a novel feature, MASK-lhs.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, Augest 13 - 15, 2021.

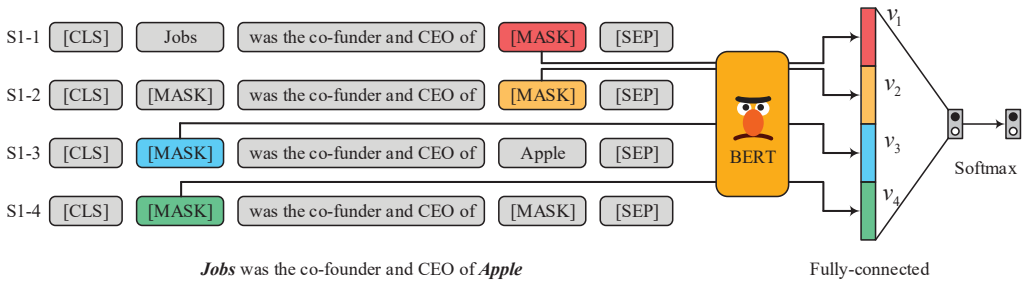

Figure 4: The training process of our noise reducers. This figure shows the MASK-lhs feature of our training process with "**Jobs** was the co-founder and CEO of **Apple**".

Algorithm 1 Overview of NS-Hunter

Input : original training set *train*, BERT-base with its 110M parameters *B*, sentence nums *n*
Output: noise reducer for each class NS_i, relation classifier *RC*

split *train* into *N*+1 classes as $train_i$ according to the label
for *i* = 1,2...*N*, **do**
    divide entity pairs in $train_i$ into source entities and target entities according to the method in Section 3.1;
    filter out the sentences whose target entity exists in the Bert-vocabulary from $train_i$ as t_i;
    sort t_i according to the dependency feature;
    In t_i, the first *n* sentences are selected as positive examples, and the last *n* sentences are taken as negative examples to form a denoise dataset nt_i;
    training NS_i based *B* with nt_i;
    apply NS_i to denoise $train_i$, remove noise-sentences, get a pure dataset of class *i* pt_i;
training *RC* based *B* with $train_0$ and $\sum_{j=1}^{N} pt_j$;

In Figure 4, we use the non-noise sentence "Jobs was the co-founder and CEO of Apple" to illustrate our MASK-lhs. First, we construct S1-1 by masking the entity 'Apple' with [MASK] (record as MASK-1), and construct S1-2 by deleting the entity 'Jobs' and masking the entity 'Apple' with [MASK] (record as MASK-2). Then v_1 is used to represent the last hidden states of [MASK-1], and v_2 is used to represent the last hidden states of [MASK-2]. From the semantic point of view, v_1 is closer to 'Apple', while v_2 is farther from 'Apple', so we think that v_1 is very different from v_2. In contrast, in noise sentence like 'During Jobs' tenure, Apple released four iPhones.', whether 'Jobs' is deleted or not, BERT can predict 'Apple', so we think that v_1 and v_2 in this noise sentence should be more similar. We expect the model to capture this semantic feature to reduce noise.

To avoid missing some information, we perform the opposite operation on two entities to construct S1-3 and S1-4(in Figure 4). Signing the last hidden states of [MASK] in the two sentences as v_3, v_4. We concatenate v_1, v_2, v_3, v_4 and add a fully connected layer. In our noise reducer, both one-token entiey and multi-token entity are replace by [MASK], so our noise reducer can deal with a sentence whether its entity is a one-token word or a multi-token word.

After each noise reducer is trained, all the sentences in the original training set will be used as the test set, and the noise sentences will be found out and eliminated. After denoising of all classes, we get a new training set. We use the data pre-processing method in MTB (Soares et al., 2019), that is, to mark the position of two entities with special symbols. For example, S-1 in Figure 1 is transformed into "#Jobs# was the co-founder and CEO of $Apple$". The steps of denoising and RC are shown in Algorithm 1.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109–1120, Hohhot, China, August 13 – 15, 2021.

4 Experments

4.1 Dataset and Evaluation

We evaluate our NS-Hunter on a widely-used public dataset NYT, which is a news corpus sampled from 294k 1989-2007 New York Times news articles (Mintz et al., 2009). Most previous works commonly generate their test sets by DS method. Such a test set can only provide an approximate measure because there are many of noise sentences in it. In contrast, Jia et al. (Jia et al., 2019) published a complete dataset ARNOR 2.0.0 [1] on the basis of the one released by Ren et al. (Ren et al., 2017) including a training set, develop set, test set and denoising dataset, in which the develop set, test set and denoising dataset are manually labeled. Jia et al. (Jia et al., 2019) removed some of the relation types which are overlapping and ambiguous or are too noisy to obtain a non-noise test sample. ARNOR 2.0.0 is the largest and most accurate dataset of sentence-level annotation at present. The denoising dataset could detect whether the model can recognize the noise sentence. We evaluate NS-Hunter on sentence-level (or instance-level) through this dataset and the details of this dataset are shown in Table 1 and Table 2.

NYT	Training	Dev	Test
#Instances	353,650	4567	4484
#Postive Instances	92707	975	1050

Table 1: Statistics of the dataset in our experiments.

NYT	Training	Dev	Test
location/location/contains	51766	479	611
business/person/company	5595	113	105
people/person/place_lived	7197	198	185
people/person/nationality	8079	117	91
people/person/place_of_birth	3173	15	13
people/location/place_of_death	1936	14	8
location/country/capital	7690	15	14
business/company/place_founded	412	0	4
location/location/neighborhood_of	5553	7	3
business/company/founders	800	6	10
people/person/children	506	11	6

Table 2: The 11 relation types retained by Jia et al. (Jia et al., 2019) and statistics of them.

Sentence-level RC is more friendly to reading comprehension tasks such as question answering and semantic analysis (Feng et al., 2018). Different from the commonly used bag level evaluation, sentence-level evaluation directly calculates precision, recall and F1-values for all instances except NA in the dataset (Ren et al., 2017). We think this evaluation method is more practical and suitable for a real world application.

4.2 Implementation Details

As we mentioned above, our NS-Hunter consists of three parts. In the first part (shown in Figure 3), we separate the entity pairs of each class except NA, and the results are shown in Table 3.

After grading and sorting the sentences, for Class k, we set:

$$n_k = min(150, 0.3 \times l_k) \tag{4}$$

and take the first n_k sentences as positive samples, and the last n_k as negative samples. Where l_k is the number of sentence whose target entity in BERT-vocablary in Class k. We separate 30% of $2n_k$ sentences into a development set.

[1] https://github.com/PaddlePaddle/Research/tree/master/NLP/ACL2019-ARNOR

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, Augest 13 - 15, 2021.

NYT	Head	Tail	Target
contains	location	location	head
company	person	business	tail
place_lived	people	location	tail
nationality	people	nation	tail
place_of_birth	people	location	tail
place_of_death	people	location	tail
capital	country	location	head
place_founded	business	location	tail
neighborhood_of	location	location	head
founders	company	person	head
children	people	person	head

Table 3: Target entity of 11 relation types.

Our noise reducers and relation classifier are based on BERT, the proportion of the one-token-entity datasets to the original dataset is 68169/92707 after looking up the BERT vocabulary. We use BERT-base-uncased with 110M parameters and set learning rate to 2e-5, the batchsize to 4 and utilize Adam for optimization. Generally, the denoising of each class can be completed in 8 epochs. The relation classification will be completed in 1 epoch and we test every 1000 batches. We set max sentence length to 450 and use Nvidia GeForce RTX 2080 Ti for training. The whole experiment will be finished in two hours.

4.3 Baselines

We compare NS-Hunter with several denoising baselines including CNN + RL_1 (Qin et al., 2018b), CNN + RL_2 (Feng et al., 2018), PCNN + ATT (Lin et al., 2016) and ARNOR (Jia et al., 2019). The experimental results of these baselines are all from the implementation of Jia et al. (Jia et al., 2019). In addition, we use the training data without denoising to classify the relation, so that we can see the good performance of our denoising method more intuitively.

4.4 Main Results

We compare NS-Hunter model with four denoising baselines. As shown in Table 4, NS-Hunter achieves state-of-the-art results in F1 metric. Results of baselines are from Jia et al.'s (Jia et al., 2019) implementation. Moreover, after denoising, we delete 55634 in 92403 relational sentences and significantly improve the precision without reducing the recall, which shows that NS-Hunter can effectively reduce the impact of noise sentences.

Method	Dev			Test		
	Pre.	Rec.	F1	Pre.	Rec.	F1
CNN+RL_1 (Qin et al., 2018b)	42.50	71.62	53.34	43.70	72.34	54.49
CNN+RL_2 (Feng et al., 2018)	42.69	72.56	53.75	44.54	73.40	55.44
PCNN+ATT (Lin et al., 2016)	**82.41**	34.10	48.24	**81.00**	35.50	49.37
ARNOR (Jia et al., 2019)	78.14	59.82	67.77	79.70	62.30	69.93
BERT without denoising (Devlin et al., 2018)	43.97	**77.32**	56.06	48.20	**78.85**	60.13
NS-Hunter (our model)	67.53	71.90	**69.65**	69.92	74.38	**72.08**

Table 4: Comparison of our NS-Hunter and other baselines. The first four methods are models for DS-RC.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, August 13 - 15, 2021.

4.5 Denoising

Our NS-Hunter reduces noise by capturing semantic differences between noise and non-noise sentences. In ARNOR 2.0.0 (Jia et al., 2019), there is a denoising dataset which includes 466 non-NA sentences labeled as noise and non-noise manually.The experimental results in Table 5 show that the effect of our NS-Hunter improved by 9.72% compared with ARNOR.

Denoise	Pre.	Rec.	F1
CNN+RL$_2$	41.35	**94.83**	57.59
ARNOR	72.04	74.01	73.01
NS-Hunter	**81.31**	84.19	**82.73**

Table 5: The experimental results of our NS-Hunter and two baselines on the denoise dataset.

4.6 Effects of Our MASK-lhs Feature

We illustrated in Section 1 that commonly-used [CLS] feature is not suitable for denoising DS dataset because the noise and non-noise sentences in a bag have the same entity pair. For the reason, we design a novel MASK-lhs feature (Figure 4), which can reduce noise by capturing the semantic differences between noise and non-noise sentences. In order to verify the superiority of the MASK-lhs feature, when training 11 classifiers in ARNOR dataset, we take the [CLS] feature as the baseline, and compare the denoising and RC effect of the model on the development set, test set and denoise set.

Feature	Pre.	Rec.	F1
CLS	**84.33**	67.21	74.80
MASK-lhs	81.31	**84.19**	**82.73**

Table 6: Experimental results of our MASK-lhs and commonly used CLS features.

Features		Pre.	Rec.	F1
Dev	CLS	52.88	71.23	60.70
	MASK-lhs	**67.53**	**71.90**	**69.65**
Test	CLS	56.12	73.24	63.55
	MASK-lhs	**69.92**	**74.38**	**72.08**

Table 7: Experimental results of our MASK-lhs and commonly used CLS features on the RC dataset.

The experimental results in Table 6 show the denoising effect of [CLS] feature and MASK-lhs feature on the denoise dataset. The experimental results in Table 7 show the RC effect of two features on the RC dataset. It can be seen that our MASK-lhs feature has increased by 7.29% in noise reduction and about 9% in RC task. Experiments show that our MASK-lhs feature can actually capture the semantic differences between noise and non-noise sentences and can better denoise DS-RC dataset than [CLS].

4.7 Apply NS-Hunter to CNN

In our NS-Hunter, the denoising part and the RC part are separated. The denoising method is plug and play for any other RC model sucn as CNN, and can also improve the classification effect. We train CNN-RC model with the denoised training set. The experimental results of the original training set come from ARNOR (Jia et al., 2019) and we use the same settings on the experiments of denoised training set. The comparison is shown in Table 8.

4.8 Effect of Entity on Denoising

Our NS-Hunter reduces noise by capturing semantic differences between noise and non-noise sentences. For example, although entity “New York” is very common, it is not in BERT-vocabulary, so “New York”

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, Augest 13 - 15, 2021.

CNN RC		Pre.	Rec.	F1
Dev	original	39.27	**73.80**	51.26
	denoising	**66.56**	64.51	**65.52**
Test	original	42.41	**76.64**	54.60
	denoising	**67.73**	64.95	**66.31**

Table 8: Experiment results of CNN-RC model before and after using our NS-Hunter.

Class	Pct.	F1 increase	
		Dev	Test
contains	59%	**24.31**	**21.24**
company	11%	4.76	5.32
place-lived	40%	4.28	9.19
nationality	48%	16.41	10.44

Table 9: The improvement of F1-value in four classes after denoising.

corresponds to two vectors in BERT's hidden state. However, in the process of training noise reducers, "New York" is replaced by [MASK], and we hope to obtain the semantic features of "New York" from a single vector corresponding to [MASK]. According to our method, this will bring some errors. Therefore, for a single class, the percentage of entities in the BERT-vocabulary should be related to the improvement of the F1-value after denoising.

In the NYT development set and test set we used, there are only four classes with more than 50 sentences. As we all know, the percentage of location in the BERT-vocabulary is far greater than that name of people. So, we show the impact of denoising module in these four classes respectively in Table 9 where **Pct.** is the percentage of original training set entities in each class included in the BERT-vocabulary. It can be seen from the Table 9 that the class "contains" with the largest percentage gets the best improvement after denoising, followed by the "nationality". The percentage of the class "company" is the smallest, and its improvement is the worst. This shows that our NS-Hunter can reduce noise according to the design principle.

5 Conclusion

After carefully observing the RC dataset generated by the DS method, we present the two semantic features, i.e. dependency and MASK-lhs feature, and propose a BERT-based denoising model NS-Hunter and a denoising approach based on the two semantic features for DS-RC. We present the dependency feature of the entity pair and use BERT-cloze to discriminate some specific sentences with BERT-vocabulary based on the dependency feature, which has strong interpretability. For general sentences generated by the DS method, we designed a novel MASK-lhs feature to capture the semantic differences between noise and non-noise sentences for denoising. The performance of NS-Hunter is better than several other denoising baselines based on CNN, PCNN and BiLSTM. Significant improvements have been made in denoising and RC task. Our denoising method can also be easily combined with other RC methods.

Because we train noise reducers for each class, even if the knowledge is updated (such as the president of a country changes), our noise reducers are still robust according to the existing similar relation patterns in the training set. However, our model may not perform well in some professional domain such as biological, because BERT generates language representation from general corpus and lacks domain-specific knowledge (Liu et al., 2019a). It may get better results if we continue to pre-training BERT in a large scale of professional texts before applying our denoising model (Gururangan et al., 2020).

Acknowledgements

The work was supported by the National Key R&D Program of China under grant 2018YFB1004700 and National Natural Science Foundation of China (61772122,61872074).

References

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019. Combining distant and direct supervision for neural relation extraction. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1858–1867. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does bert solve commonsense task via commonsense knowledge?

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5779–5786. AAAI Press.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distant supervision for relation extraction via instance-level adversarial training. *CoRR*, abs/1805.10959.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3060–3066. AAAI Press.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: attention regularization based noise reduction for distant supervision relation classification. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1399–1408. Association for Computational Linguistics.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Hai Zhao. 2019a. GAN driven semi-distant supervision for relation extraction. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3026–3035. Association for Computational Linguistics.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2019b. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. *CoRR*, abs/1911.11899.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1790–1795. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, Augest 13 - 15, 2021.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, and Janyce Wiebe, editors, *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

Maria Pershina, Bonan Min, Wei Xu, and Ralph Grishman. 2014. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 732–738. The Association for Computer Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. DSGAN: generative adversarial training for distant supervision relation extraction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 496–505. Association for Computational Linguistics.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2137–2147. Association for Computational Linguistics.

Jianfeng Qu, Wen Hua, Dantong Ouyang, Xiaofang Zhou, and Ximing Li. 2019. A fine-grained and noise-aware method for neural relation extraction. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 659–668. ACM.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1015–1024. ACM.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2895–2905. Association for Computational Linguistics.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2019. A novel cascade binary tagging framework for relational triple extraction.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2361–2364. ACM.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In Jan Hajic and Junichi Tsujii, editors, *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344. ACL.

Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large scaled relation extraction with reinforcement learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5658–5665. AAAI Press.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1109-1120, Hohhot, China, Augest 13 - 15, 2021.