A Zero-Shot Relation Extraction Approach Based on Contrast Learning

Hongyu Zhu, Jun Zeng, Yu Yang, Yingbo Wu
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
zengjun@cqu.edu.cn

Abstract—The most significant advantage of the unseen relation extraction is that it can recognize unlabeled relations. While Zero-Shot Learning can meet the requirements of the identification of unseen relation through relation description information without labeled datasets. However, unseen relation extraction requires an effective method in representation and generalization, which become a challenge for zero-shot learning approach. In this paper, we propose a Zero-Shot learning Relation Extraction based on Contrastive learning Model (ZRCM) to capture deep interrelation text information. We design a comparison sample generation method which can produce several instances for one input sentence and compare the distance between positive instance and negative ones, so as to improve the hidden text information mining ability. Experiments conducted on relation extraction common datasets confirmed the promotion of ZRCM compared with the existing methods. Especially, our model can improve the F1 value by up to 7% at best. When there are fewer unseen relations to predict, our model can achieve better performance.

Index Terms—relation extraction, zero-shot learning, contrastive learning

I. Introduction

Relation Extraction (RE) is a task of extracting the possible relation between a given ordered pair of entities and the relevant context information from one passage or sentence. Relation extraction is a predecessor task of NLP and plays an important role in many downstream tasks, such as constructing or expanding knowledge graph [15] and question answering system [8]. Existing methods for RE often require large-scale labeled datasets to conduct the training process, these labeled datasets require manual pre-process, which are time-consuming and labor-intensive [13]. One possible way to solve this problem is to use distant supervision generate annotation datasets [10]. However, existing datasets are always difficult to cover all the relations, due to the variety of relations. If the training datasets do not contain the labels for all relations, the existing supervised learning methods for relation extraction cannot handle this situation adequately.

It is unreasonable to assume that training data always contains relations which need to be distinguish in real world. Therefore, it is crucial to explore new models to predict new classes which are not defined or observed in advance, such tasks are called Zero-Shot learning (ZSL) [7]. By applying ZSL to relation extraction, Zero-Shot Relation Extraction (ZS-RE) become a paradigm which has a logical strategy in dealing with unseen relation. ZS-RE can identify new relations which don't have labeled data. In other words, it requires the model to predict whether an input sentence containing with two entities matches one relation whose description can be found from a series of relations. The core of ZS-RE is to find the indicative text information that can be used to judge whether the input sentence contains unseen relations, and the model can use it as a comparison basis to define the type of the relation among the input sentence. ZS-RE can also be regarded as a textual entailment task [17] in a broad sense, which is applied to identify new relations without corresponding labeled data for training, or requires the model to predicts whether the input sentence containing two entities also contains a relation matching the description of a given relation. The unseen relations are predicted through the existing labeled data, thus avoiding the over-dependence on the training data of a large number of sequence annotation tasks.

Studys on ZS-RE can be divided into two categories by the indicative text information of the comparison part. One is the generative method. The generative method means that the index text information required by ZSL need to be generated through the existing model. For example, [14] question the relations in the sentences using the generation method of Query in Q&A and take these questions as the key information for subsequent model training. Another approach we called self-labeling method uses explanatory text information in the network as indicator text information. Through the contrastive learning method [3], [4], the self-labeling method can achieve the purpose of training model by reducing the distance between the vector representation of input sentences generated by the representation learning method and the indicator text information. We call it self-labeling method because the text information indicator we need is usually the definition of a word or phrase, such as naturally marked and interpretable texts, and these texts are often highly reliable and relatively easy to obtain, such as Wiki-data. Owing to the advantages of the self-labeling

DOI reference number: 10.18293/SEKE2022-032

method, this paper adopts the relational description as the indicative text information of the contrastive learning, and identifies the unseen relations by extending the semantic space with generalization ability through the representation learning. But it is difficult to learn the semantic space with robustness and generalization, especially the over-fitting problem in ZSL and the uncertainty of unseen relation prediction.

In order to improve the performance of the model and get more generalization ability of the model and weaken the influence of the fitting problem, we propose an adversarial contrastive learning method for relation extraction, design a negative sample generator for zero-shot contrastive learning to generate different negative complements which can complete missing information or weaken the information that positive samples pay much attention to, so that the model can obtain better result in generalization ability. Finally, the model parameters are optimized by loss function specially designed to get better performance. This paper provides a method to improve the accuracy and stability of ZSL relation extraction by constructing a targeted contrastive learning method based on the comparative information of relation descriptions and BERT [6] pre-training language model. The contribution of this article can be summarized as follows:

- In general, we use relational description information to construct a contrastive training method for zero-shot relation extraction, which not only has the interpretability and simplicity of natural language, but also has significant advantages for the relation extraction task involving unseen relations.
- Methodologically, considering the generalization ability of the model, we construct a generation method of contrastive learning instances for relation extraction, and a negative samples generator for adversarial training, which lead to the significant improvement in the model accuracy, recall and F1 value.
- Experimentally, sufficient experiments based on two representative datasets demonstrate the effectiveness of our contrastive learning method for relation extraction and the effectiveness of the negative sample generation method in contrastive learning.

II. Problem Definition

We consult the definition of zero-shot relation extraction given by [1] and transform it to suit our model requires. There is a text collection W, which includes a number of sentences S, each sentence contains two entities e_1 and e_2 respectively. Besides, each sentence also contains a specific relation R with its description d for the entity pair e_1 and e_2. There is another text collection $\bar{W}$, together with a serious of sentences $\bar{S}$, each of which also contains an entity pair $\bar{e_1}$ and $\bar{e_2}$. But we do not know what the relation $\bar{R}$ between entities is. Our goal is to train a model M, which deduce the relation $\bar{R}$ in sentences $\bar{S}$ from training by input sentences S together with its relation description d. That is to say, we use model to infer unseen relation with labeled data. Out input can be expressed as $P = (S_0, e_1, e_2, R, d)$, through the model, the result can be indicated as $M(P) \Rightarrow \bar{R} \in \bar{S}$. That is the process of judging unseen relation by seen relation.

In the process of model training, in order to capture more detailed textual information, we put forward a Negative Sample Generator. It generates multiple different representation instances of the same sentence, these negative samples S_1, S_2, S_3 carry different levels of information containing the original data. We improve the model effect by reasonably adjusting the composition and training of examples to maximize the span between positive and negative samples. Then, our method is updated as: $M(\bar{P}) \Rightarrow \bar{R} \in \bar{S}, \bar{P} = (S_0, e_1, e_2, R, d)$, where $i = 0, 1, 2, 3$. $\bar{P}$ represents the new input to the model. This approach promotes our method to mine more detailed hierarchical information, as detailed in chapter 3.

Relevant researches on zero-shot relation extraction are few. This paper is similar to [1], who use training instances formed by input sentences and relation descriptions map into the embedding space by minimizing the distance between sentences and relation descriptions jointly and classify the seen and unseen relations. Since zero-shot relation extraction can be regarded as a text implication task, methods containing text implication thought can also achieve similar effects, like Enhanced Sequential Inference Model (ESIM) [2] and Conditioned Inference Model (CIM) [16]. By pairing each input sentence with each relation description, they trained the model to answer whether the paired text was contradictory or implicit. These models can infer and predict unseen relations by training input sentences and unseen relation descriptions. Contrastive Learning is often regarded as unsupervised learning or self-supervised learning method. Its core purpose is to obtain vector representations which are compatible with downstream work by limited data and labels. Thus, the use of labeled data and the choice of training methods are very important. In this paper, contrastive learning is included in the comparison of text-implied task methods, and the learning ability and generalization ability of model representation learning are improved by increasing the available information with samples generated.

III. The Proposed Model

In this section we discuss our model ZRCM whose basic process is shown in 1 The Negative Sample Generator and our training method are also explained in detail.

A. Model Process

We tokenize the sentences and send them into encoder to obtain contextual representation. For sentence $S_0^i, S_0^i \in S_0$, every token is represented as a vector H_i, where $i \in [0, n]$ and n denotes the number of tokens in S_0^i. [CLS] contain the whole information in sentence indicated as H_0,

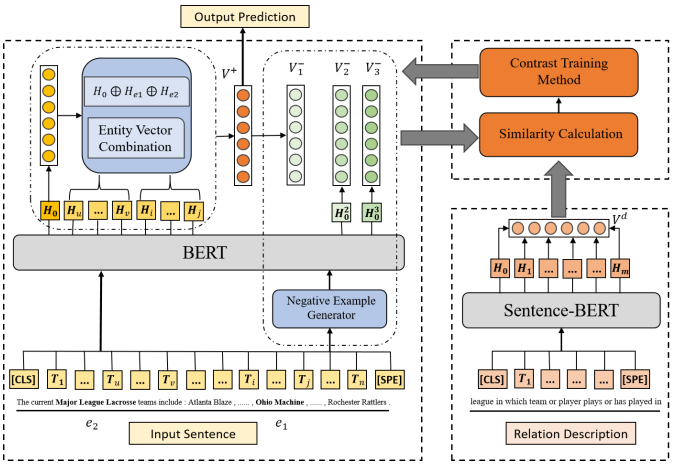

Fig. 1. The architecture of ZRCM.

where [CLS] is the classification token for sentences . We extract the representation of two entities and concatenate the vector representations respectively:

$$H_{e1} = \frac{1}{j-i+1}\sum_{a=i}^{j} H_a, H_a \in H_1^e, \tag{1}$$

$$H_{e2} = \frac{1}{v-u+1}\sum_{a=u}^{v} H_a, H_a \in H_2^e, \tag{2}$$

where i and j represent the start and end position of the tokens for e_1, u and v represent the start and end position of the tokens for e_2. $H_1^e = \{H_a | i \leq a \leq j\}, H_2^e = \{H_b | u \leq b \leq v\}$. The vector representing the entities are finally obtained by mean pooling, respectively. To capture the further information in entities, we concatenate H_{e1}, H_{e2} and H_0 by a fully connected layer and activation operation which are added to H_0: $\widetilde{V}^+ = H_0 + W(tanh([H_0 \bigoplus H_{e1} \bigoplus H_{e2}])) + b$, where W and b are the parameters that model needs to learn, $\bigoplus$ represents the vector concatenation. Then $\widetilde{V}^+$ is sent through a fully connected layer and activation to get V^+, which is one of the most important generate representation. It is clear that the entity pairs confirmed as the kernel of RE, but we do not want to pay too much attention to it to weaken other important information in the input sentence. Thus, we use residual network structure concatenate H_{e1}, H_{e2} and H_0 to construct V^+, which represents the relation representation contained in our input sentence for this model. On the other hand, the relation description d in input is feed to Sentence-BERT model, and which also be represented as a vector representation, donated as V^d. Noted that we need relation descriptions to modify our input representation, which are fixed and generated by the Sentence-BERT provided by [12]. The closer the distance $D(V^+, V^d)$ between V^+ and V^d, the more expressive V^+ the model obtain, where D represents a distance algorithm for calculating similarity of distance, including three possible choices of Euclidean distance, inner product and cosine similarity. In order to fully dig out the information of the potential implication relation contained of the input sentences we generate a negative sample generator for the zero-shot relation extraction. The detailed information of the negative sample generator will be introduced in chapter 3. We send input sentence S_0^i to negative sample generator and we will get three different processed sentences S_1^i, S_2^i, S_3^i, which will be feed into our model together with S_0^i in our input $\bar{P}$.

B. Negative Sample Generator

The choice of negative samples has a considerable influence on the effect of comparative learning. Negative samples are generally borne by other positive samples in batch. Although this method is simple and convenient, it obviously has certain shortcomings. From an empirical point of view, the negative samples selected in this way are highly random and may cause fluctuations in the training of the model, and the validity of negative samples produced in this way needs further confirmation. Such selections of negative samples are difficult to stand up in terms of interpretability. The false negatives generated by other sentences in the batch may have a similar relation description with the positive sample, which will greatly interfere with the training of the model. Therefore, we design a negative sample generator for zero-shot relation extraction. It can generate three different types of negative samples from the positive samples. We named them Random Negative Samples (RNS), Relational Negative Samples (ReNS), and Entity Negative Samples (ENS). What needs to be mentioned is that relational negative samples and entity negative samples are weak negative samples produced in order to cooperate with the contrastive learning of positive samples. They have a small gap with the positive samples and need to be used in conjunction.

For a positive sample of a trained sentence S_0^i, we take one vector representation from other sentences in the same batch which is the farthest away from the vector representation of this relation as the random negative sample S_1^i. In order to retain the real information and produce a large gap with the vector representation of the positive sample, we choose to use the same method as the positive sample for training and generate the corresponding vector representation V_1^-. Then we can express it as: $D(V_1^-, V^d) = max(D(V_i^+, V^d)), i \in b$, where b represents the size of the batch size, V_i^+ represents other sentences contained in the same batch, and D represents the similarity function. Although the selection of random negative samples is uncertainty, it can improve the generalization ability of the model through randomness in zero-shot learning scenarios.

In order to more deeply capture the relational information contained in the sentence, we have generated weak negative samples V_2^-, that is, relational negative samples.V_2^- is generated by mask tokens that may directly

indicate relational words in the sentence. We maximize the distance $D(V_2^-, V^d)$ by the sentence and the vector representation of the relation description V^d to obtain information which may be missed by the positive sample as a supplement. Many sentences may not contain tokens which can directly indicate relation words. When this happens, we use a certain percentage of tokens in the random mask sentence as an alternative method (the entity pair tokens are not included).

For zero-shot tasks, over-fitting is one of the most serious problems. We think this is a major problem that limits its generalization performance for this model that pays too much attention to entity information. Therefore, we mask the tokens of the entity pair in the sentence to generate the negative input of the entity, and maximize the distance $D(V_3^-, V^d)$ between the generated negative vector representation V_3^- and the relation representation V^d.

For the last two methods, they generate weak negative samples and have no need to focus on the entities. We directly use the hidden layer corresponding to [CLS] to pass through an activation function layer and a dropout layer as V_2^- and V_3^-, respectively. In this way, through the negative sample generator, a positive sample generates several negative samples with different emphasis information.

C. Training

The training of ZRCM consists of two objectives. Firstly, by generating suitable positive and negative samples, and increasing the span between positive and negative samples as much as possible to obtain a more generalized model effect. We compare the distance of the positive case with the three negative case distances as a pair of calculations:

$$C(D(V^+, V^d), D(V_i^-, V^d))$$

$$= max(0, \gamma - D(V^+, V^d) + D(V_i^-, V^d)), \tag{3}$$

where $i = 1, 2, 3$. Then the goal of our model is expressed as:

$$\mathcal{C} = \sum_{i}^{1,2,3} [C(D(V^+, V^d), D(V_i^-, V^d))] \tag{4}$$

where γ is a hyper-parameter, whose purpose is to keep a certain buffer space for the distance difference between the positive sample and the negative sample. In the training, we iterate the computation to make $D(V^+, V^d)$ obtain a larger value while ensuring that $D(V_i^-, V^d)$ is smaller, that is we increase the distance between positive samples and the relation discription representations while reduce the distance between negative samples and the relation discription representations.

Our second objective is to use cross-entropy loss to maximize the accuracy of Relation Label Classification based on visible relations:

$$\mathcal{D} = max(0, \sum_{j}^{0,2,3} (-1)^{2\times j} V_j^R log(\bar{V_j^R}) + \beta), \tag{5}$$

where V_j^R represents the visible relation, $\bar{V_j^R}$ represent its relation represents the probability distribution of the corresponding visible relation prediction, and β is a hyper-parameter, in order to ensure the full use of the training data. In addition, we also generated the corresponding visible relation prediction distributions for the other negative samples which are divided and can ensure the sufficiency of the negative samples we generate when used in the first objective. Because the syntax structure of the input two negative samples is very similar to the positive sample, and it is necessary to ensure the difference in this way. The larger the relation prediction gaps are, the more representative the negative samples we generated and the more helpful for generalization ability. In general, the larger the $\mathcal{D}$, the higher the probability that the predictions are correct.

Combining the two objectives described above, our final objective function can be expressed as:

$$L = (1 - \alpha) \times \mathcal{C} - \alpha \times \mathcal{D} \tag{6}$$

where $\mathcal{C}$ comes from Eq.(4), and $\mathcal{D}$ comes from Eq.(5), which are the hyper-parameters. All the hyper-parameters mentioned in the model will be studied and discussed in detail in the subsequent experimental part.

IV. Experiments

A. Experimental Setup

Datasets. We use two datasets for experiments, FewRel [9] and Wiki-ZSL [5]. FewRel has 70,000 sentences selected by a large number of crowd workers from Wikipedia, which contains about 100 relations. Then use the distant supervision method to complete the preliminary labeling, and then manually filter out the wrong sentences, so that the dataset becomes a clean RC dataset, which has 56,000 examples containing 80 different relationships. For another data set Wiki-ZSL, which originally came from Wiki-KB, it was also generated with a distant supervision method, with 93483 examples and 113 different relations. The two data sets have one thing in common, that is, they are both built on the basis of data in Wikipedia, which allows them to be accurately linked to the Wiki-data knowledge base. This provides possibility and great convenience for the zero-shot relation extraction.

ZSL Experimental Setup. We divide a data set into three different predictive relation quantities $\mathfrak{a}$. That is, one part of the data set is used for training, and the other part is used for prediction. $\mathfrak{a}$ has three possible options: 5, 10, and 15. In order to meet the requirements of zero-shot learning, it is necessary to ensure that there

is no intersection between the trained relational data and the predicted relational data. We use Precision, Recall and F1 value as a measurement method to evaluate the effect of the experiment. We repeat the experiment for more than 5 times, randomly select relations as test set, make the rest of them as training set, report the best result of every single experiment and evaluate the final results comprehensively. We do our best to ensure the comparability of our experimental method and other comparison methods.

Comparison Methods. R-BERT is a supervised method of relation extraction. It has excellent results in fully supervised relation extraction experiments but performs poorly on zero-shot prediction tasks. ESIM [2] and CIM [16] are two texts which contain tasks. They accept sentences and relation descriptions as input, and output a binary label indicating whether they match semantically. ZS-BERT is the baseline of this experiment, and it has an experimental effect far superior to other models by using zero-shot learning. The experimental results of other comparison methods are from [1]. In general, we hope to show the advantages or disadvantages of this method by comparing the results with other methods.

Parameter Settings. Our model is based on Hugging Face and PyTorch. We use the Adam [11] optimizer, the batch size is set to 4, the size of the hidden layer is 768, the embedding dimensions of the input sentences and the dimensions of the attribute vectors are 1024. To make the experiments easy to compare, we used exactly the same data set with the traditional evaluation indicators: Precision, Recall and F1. For different datasets our hyperparameters are not the same. For FewRel, $\alpha = 0.4, \beta = 4, \gamma = 7.5$, for Wiki-ZSL, $\alpha = 0.4, \beta = 0, \gamma = 7.5$. α is the weight parameter of the balances loss function. The similarity function D has been compared through many experiments and found that using inner product will achieve a more stable and high-quality effect.

TABLE I
The comparative results of the experiments

	Wiki-ZSL			FewRel		
	$\mathfrak{a} = 5$			$\mathfrak{a} = 5$		
	P	R	F1	P	R	F1
R-BERT	39.22	43.27	41.15	42.19	48.61	45.17
ESIM	48.58	47.74	48.16	56.27	58.44	57.33
CIM	49.63	48.81	49.22	58.05	61.92	59.92
ZS-BERT	71.54	72.39	71.96	76.96	78.86	77.90
ZRCM	76.15	77.1	76.6	86.70	84.51	85.60
	$\mathfrak{a} = 10$			$\mathfrak{a} = 10$		
	P	R	F1	P	R	F1
R-BERT	26.18	29.69	27.82	25.52	33.02	28.20
ESIM	44.12	45.46	44.78	42.89	44.17	43.52
CIM	46.54	47.90	45.57	47.39	49.11	48.23
ZS-BERT	60.51	60.98	60.74	56.92	57.59	57.25
ZRCM	62.41	64.16	63.27	53.67	53.96	53.81
	$\mathfrak{a} = 15$			$\mathfrak{a} = 15$		
	P	R	F1	P	R	F1
R-BERT	17.31	18.82	18.03	16.95	19.37	18.08
ESIM	27.31	29.62	28.42	29.15	31.59	30.32
CIM	29.17	30.58	29.86	31.83	33.06	32.43
ZS-BERT	34.12	34.38	34.25	35.54	38.19	36.82
ZRCM	33.47	36.71	35.01	40.27	40.72	40.50

B. Experimental analysis

Main Experiment. We predict the experimental results of different numbers of unseen relations which can be seen from Table 1. First of all, we can see that our model's effect is significantly better than other models, especially when $\mathfrak{a} = 5$. Our model outperforms the second model about 7.7% at F1 value in FewRel dataset which can show the ability of case comparison method to capture potential information. For $\mathfrak{a} = 15$ our model can also achieve F1 value increase of about 3% at best, which reflects our model's superiority in predicting more generalization ability of unseen relations. When $\mathfrak{a} = 10$ ZRCM lags ZS-BERT by about 3.4% in FewRel, we believe that it may be due to insufficient data that the negative samples and positive samples are not reasonably divided. On the whole, when $\mathfrak{a}$ is smaller, the predicted unseen relations are fewer and the experimental precision recall and F1 value will be ignificantly higher than when $\mathfrak{a}$ is larger. This is predictable. On the one hand, we only need to predict fewer target relations, and at the same time, the sources of information we can obtain are also increasing. It can be seen that although the textual implication models such as ESIM and CIM have a higher improvement than R-BERT, there is a big gap between ZS-BERT and this experiment, which shows that the textual implication tasks cannot be perfectly covered and fit unseen. For ZS-BERT and ZRCM, it can be seen that the results of our model have substantial improvements on ZS-BERT, which reflects the effectiveness and superiority of the overall process setting of our model.

Ablation. In order to fully demonstrate the effectiveness of our negative samples, we design the ablation experiments for them based on the method of controlling variables. That is to say, we eliminate the distance constraint between the label and the input hidden layer vector and only consider the unilateral impact of negative samples on the experimental result. ZRCM is our model, ZRCM* expresses our first goal, that is, the result of ZRCM after removing the Relation Label Classification. RNS, ReNS and ENS are the samples that we build with Negative Sample Generator to compare with the positive samples. It can be obtained from the data analysis in *Table*2 that ZRCM get the better results than ZRCM*, which proves

the contribution of our goal 1 to the overall model effect. Besides, for three comparative negative samples we can see that RNE contributes the most to the result which is understandable, because although other comparative negative samples may carry more deep attributes, they still need a sufficient distance to optimize the model. Since different data sets have different text features, it is reasonable that the results of the same comparison negative sample on the two data sets are different.

TABLE II
The impact of different negative sampling methods

F1	FewRel	Wiki-ZSL
ZRCM*	0.4387	0.2950
ZRCM* - RNS	0.1314	0.0748
ZRCM* - ReNS	0.4178	0.3236
ZRCM* - ENS	0.4591	0.3098
ZRCM	0.4544	0.3298

Hyperparameter Experiment. In order to achieve the optimal performance of our model, we conduct extensive experiments to judge the effect of different hyperparameters on the performance.

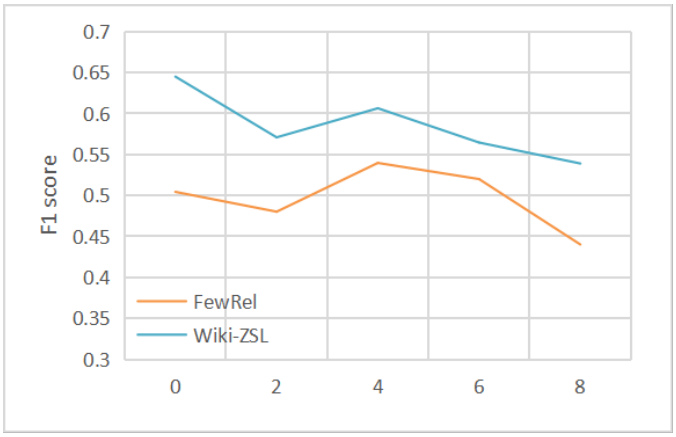

Fig. 2. Changes in F1 under the influence of β

In particular, we show how changes in the parameter β in Eq.(5) affect the model performance, β is a boundary parameter. To determine that the objective 2 of the experiment obtains a more ideal optimization distance, we tried different values of β when $\mathfrak{a} = 10$. For generating better negative samples, objective 2 is designed to help the model generate more representative negative representations, which makes the vector representation of the labels have greater distance from the vector representation of the negative samples generated by the negative sample generator. The results on two datasets are exhibited on Fig. 2, the two curves have the same trend, but they achieve the best performance in different place. It's reasonable for the inherent differences in the textual information of the two datasets.

V. Conclusion

In this work, we present a novel method matching representation learning for Zero-Shot Relation Extraction. With the Negative Sample Generator, our model can capture depth information for the input sentences. Besides, we use multi-task learning structure with negative sample training model. Results show that our model can substantially improve the performance, we also carry out extensive experiments which can verify the effectiveness of the designed adversarial training. The ability to understand and summarize the text and the problem of over-fitting are the important factors which limit the effectiveness of the model, which also forms the basic idea of our method.

Acknowledgment

This research is supported by the National Key Research and Development Program of China (Grant No. 2019YFB1706101), Natural Science Foundation of Chongqing, China (No. cstc2020jcyj-msxmX0900), and the Funda-mental Research Funds for the Central Universities (Project No. 2020CDJ-LHZZ-040)

References

[1] Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards Zero-Shot Relation Extraction with Attribute Representation Learning. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Pages: 3470–3479. Association for Computational Linguistics.

[2] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1657–1668 Vancouver, Canada. Association for Computational Linguistics.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119:1597-1607.

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, Kaiming He. 2020. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

[5] Sorokin Daniil, and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1784-1789 Copenhagen, Denmark. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[7] Kodirov Elyor, Tao Xiang, and Shaogang Gong. 2017. Semantic autoencoder for zero-shot learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3174-3183.

[8] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, Jiwei Li. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1340-1350 Florence, Italy. Association for Computational Linguistics.

[9] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4803-4809 Brussels, Belgium. Association for Computational Linguistics.

[10] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, page 3060–3066. AAAI Press.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980.

[12] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982-3992, Hong Kong, China. Association for Computational Linguistics.

[13] Yong Shi, Yang Xiao, Pei Quan, Minglong Lei, Lingfeng Niu. 2021. Distant Supervision Relation Extraction via Adaptive dependency-path and Additional Knowledge Graph Supervision. Neural Networks, 134:42-53.

[14] Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, Eneko Agirre. 2021. Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1199–1212. Association for Computational Linguistics.

[15] Haoze Yu, Haisheng Li, Dianhui Mao & Qiang Cai. 2020. A relationship extraction method for domain knowledge graph construction. World Wide Web 23.2 (2020): 735-753.

[16] Shu Zhang, Dequan Zheng, Xinchen Hu and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. Proceedings of the 29th Pacific Asia conference on language, information and computation, pages 73-78 Shanghai, China.

[17] Obamuyide, Abiola, and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER).