Improving Distant Supervised Relation Extraction via Jointly Training on Instances

Binjun Zhu[1], Yijie Zhang[2], Chao Wang[1], Changjian Hu[1], and Feiyu Xu[1]

[1] Lenovo Research, No. 10, East Xibeiwang Rd., Haidian District, Beijing, China
{zhubj4,wangchao31,hucj1,fxu}@lenovo.com
[2] Dept. of Computer Science and Technology, Tsinghua University, Beijing, China
yj-zhang15@mails.tsinghua.edu.cn

Abstract. In this paper, we present the winning model for Bag Track in Inter-Personal Relation Extraction. We incorporate BERT, a large pre-trained language model with multi-instance learning for bag-level relation extraction. To further take advantage of the large pre-trained language model as the most advanced language understanding tool, we propose a multi-task learning scheme that learns bag-level representation together with sentence-level representation. Our results demonstrate that the auxiliary task of sentence-level prediction significantly benefits the model performance for bag-level prediction.

Keywords: Relation Extraction · Distant Supervision · Jointly Training

1 Introduction

The task of relation extraction (RE) is important in natural language processing, and is quite distinct between sentence-level RE and bag-label RE. Sentence-level relation extraction is usually a fully supervised task aiming to predict the relation between entities given a sentence describing them. For example, given the entity pair <Bob, Alice> and the sentence "Bob, the son of Alice, was born in the United States", we can determine an inter-personal relation between them that Bob is the son of Alice.

Bag-level relation extraction is more complicated for two reasons: first, data is derived by distant supervision[8]. Distant supervision automatically crafts the dataset through assuming all collected sentences with the same entity pair have the same relation, which inevitably introduces wrong labeled data. Second, a bag-level prediction needs the utilization of all information in one bag. While large pre-trained language models have shown their impressive ability in capturing semantic cues[3][9][10][19], we still need to explore ways of combining sentence representations in a bag into single bag-level representation.

For sentence-level relation extraction, it is challenging to design a model under distant supervision, since there is too much noise if we follow a regular supervised training scheme. [4] proposed a reinforcement learning method to learn a sentence selector for data denoising. For bag-level relation extraction,

previous work used multi-instance learning framework to deal with bag-level relation extraction under distant supervision[20][6][7]. However, they didn't adopt the state-of-the-art technology in natural language understanding, i.e. large pre-trained language models.

In this paper, we present the winning system of Bag-Track in Inter-Personal Relation Extraction (IPRE)[16] challenge. IPRE dataset is constructed under distant supervision with 34 types of inter-personal relations. The task of Sent-Track is to build a model for sentence-level prediction, and the task of Bag-Track is to distinguish all relations mentioned in a bag. The main challenge is the data noise from distant supervision. We experiment sentence-level classification by using BERT[3] as the sentence encoder, then apply multi-instance learning framework to overcome the problem of noisy data. Since BERT output for each sentence contains rich semantic information, we further propose a multi-task learning scheme for bag-level classification to combine the advantages of both. Auxiliary task and multi-task learning have been successful in the field of natural language processing[11][3][10]. Just like BERT which employs both masked token prediction and next sentence prediction in its training process, we enhance our bag representation through training with both sentence-level representation and bag-level representation.

Our main contributions are listed as follows: (1) Explore several approaches for ensemble and feature engineering to enhance model performance. (2) Incorporate large scale pre-trained language model with multi-instance learning. (3) Propose a multi-task framework for jointly training with both bag-level representation and sentence-level representation.

2 Related Work

Relation extraction from unstructured text data is a fundamental problem in natural language processing and is of great benefit to the construction of large scale knowledge graph. Typical neural network models have achieved significant success in relation extraction for clean data[20][13][17]. Recently proposed large pre-trained language models [3][9][10][19] are also promising in relation extraction. Several attempts that apply BERT to relation extraction have shown that, without further constructing sophisticated neural networks, models based on BERT perform comparable or better results than typical CNN or LSTM methods[14][21].

One critical challenge for developing practical relation extraction models is that building a large scale supervised dataset is expensive and time-consuming. The alternative paradigm to build training dataset is distant supervision[8], which automatically labels sentences from large unsupervised corpus if they contain the same entity pair that exists in Freebase, alleviates human effort for building large scale labeled dataset. However, models trained under distant supervision often suffer from noisy labeling problem[4]. Multi-instance learning(MIL)[12][15][20] was proposed to solve the wrong label issue by assuming that each bag has at least one sentence indicates the relation of its label. While

at-least-one MIL can avoid training with large amount of wrong labeled data, it also loses possibly informative data in bags that contains multiple correct labeled sentences. Other bag-level denoising methods such as instance-level attention mechanism[6], soft-label[7] and cross-max[5] incorporate all sentences in a bag as the evidence of classification.

While bag-level models with multi-instance learning addresses noisy labeling problem and perform better results for bag-level prediction than sentence-level models, they only use summarized representations of sentences in a bag in the training process, while the original individual outputs of sentence encoder may be a better representation of semantic. [1] introduced a multi-task learning setup for relation extraction under distant supervision. However, their method requires additional supervised data and only does binery classification on sentences to benifit attention weights for bag-level denoising. In this paper, we apply BERT as the sentence encoder, and introduce sentence-level prediction as the auxiliary task, directly enhance the representing ability of BERT through training with both bag-level representation and sentence-level representation.

3 Methology

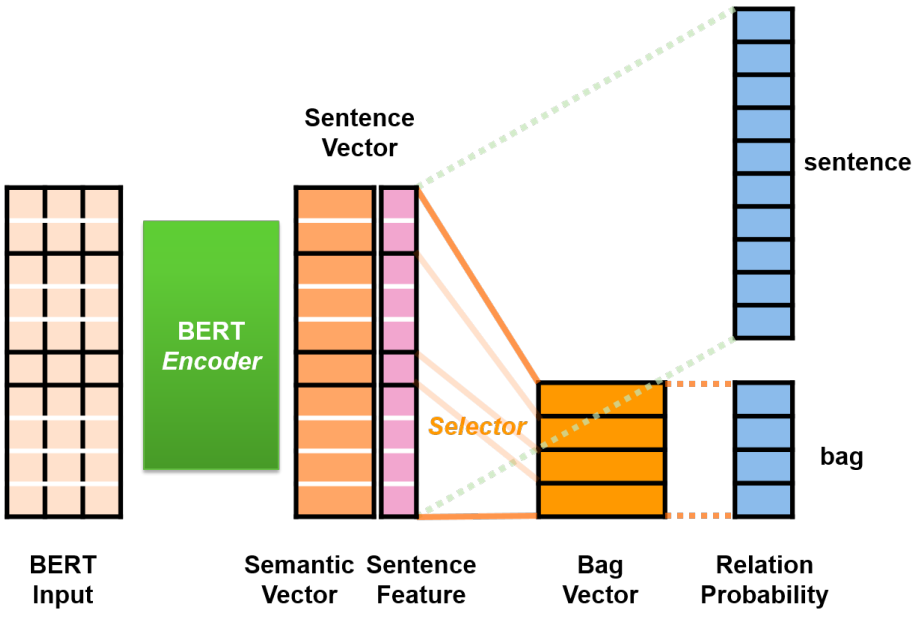

Fig. 1. The structure of multi-task relation extraction model.

3.1 Neural Architecture

As shown in Fig. 1, a bag contains multiple sentences, and all the sentences are all considered as the model input. We employ BERT as the sentence encoder and transform the input into the semantic vectors. The semantic embedding vector and manually designed sentence feature vector are concatenated to represent the sentence embedding vector. A bag-level denoising strategy is then applied to condense multiple sentence embeddings into one single bag embedding for each input bag. Fully connected layers are added to produce probability distributions for sentence relation likelihood and bag relation likelihood. Instead of training the model only to predict bag labels as the traditional MIL approach, our proposed model also minimizes cross-entropy loss for sentence-level prediction.

3.2 Data Preprocessing

The original distant supervision data (bag track) can be defined as follows:

$$ins =< e1, e2, text; label_{ins} > \tag{1}$$

$$bag =< (ins_1, ins_2, ..., ins_m); label_{bag} > \tag{2}$$

$$D_{bag} = (bag_1, bag_2, ..., bag_n) \tag{3}$$

where D_{bag} is distant supervised dataset and composed with bags. A bag is an instance collection in which all instances describe the same entity pair. The label of a bag is also a collection of all instance labels in a bag.

For simplification, the model predicts exactly one relation label for a bag, although bags may have multiple labels in some Multi-Instance Multi-Label settings. For those bags in training data that have more than one relation labels, we assign the most common positive relation as the bag label or NA if there is no positive relation.

$$Input = (ins_{1,1}, ..., ins_{1,m_1}, ins_{2,1}, ..., ins_{2,m_2}, ..., ins_{n,m_n}) \tag{4}$$

All the instances in each bag are concatenated as model input, which is defined in Equ.4. The description texts of instances are converted into token sequence with the same method in [3][18]. In particular, each Chinese character is a token, thus the number of tokens is exactly the same as the length of text.

To prevent BERT from overfitting the entity names, we replace all the given entity names $e1, e2$ in the text by two pre-defined entity names $head, tail$. On the other hand, as the original entity pair $e1, e2$ does not appear in the corresponding description text, the model can't capture the character information of name from the text. Thus we introduce manually designed features for the instance representation layer.

3.3 BERT Encoder

The pre-trained BERT model encodes the instances and also gets fine-tuned during the training process. The final hidden state corresponding to the first input token [CLS] is used as the sequence representation. Thus each instance is encoded into a semantic embedding vector $V_{sem} \in \mathbb{R}^{d_{BERT}}$, where d_{BERT} is the dimension of BERT hidden states.

3.4 Sentence Representation

To fully utilize all existent information, especially to save the loss from anonymizing entity pairs, we manually design 4 types of features, 1) Length, 2) Entity Gender, 3) Entity pair Similarity, 4) Name Style. The sentence feature is defined as Equ.5, where $Feat_{manual}$ is the manually designed feature vector, FC is fully connection layer and k is the dimension of the manual feature vector.

$$Feat_{sent} = FC(V_{sem}, Feat_{manual}), Feat_{manual} \in \mathbb{R}^k, Feat_{sent} \in \mathbb{R}^k \quad (5)$$

To represent the sentence, the semantic embedding vector V_{sem} and the sentence feature vector $Feat_{sent}$ are concatenated into the sentence embedding vector $Repre_{sent}$

$$Repre_{sent} = V_{sem} \oplus Feat_{sent} \quad (6)$$

3.5 Bag Representation

Each instance in a bag $(ins_1, ..., ins_m)$ has one sentence representation vector. To fully make use of information across sentences, many selectors are proposed to encode the sentence-level representation into bag-level representation[6][5]. Our model adopts cross-sentence max-pooling[5] strategy for encoding the sentence vectors in each bag.

Suppose that there are m sentences with the same entity pair in a bag, and p_i^j denotes the i-th component of the vector representation of the j-th sentence, cross-sentence max-pooling aggregates all sentence representations into an entity pair-level representation $g = g(g_1, g_2, ..., g_k)$, where:

$$g_i = \max(p_i^1, p_i^2, ..., p_i^m) \quad (7)$$

3.6 Joint Loss

Two fully connected layers are utilized to reduce the dimension of sentence and bag embedding vectors, and the softmax operation is applied to calculate the relation probability:

$$p_{sent} = Softmax(FC_{sent}(Repre_{sent})), p_{sent} \in \mathbb{R}^{d_{rel}}; \quad (8)$$

$$p_{bag} = Softmax(FC_{bag}(Repre_{bag})), p_{bag} \in \mathbb{R}^{d_{rel}}; \quad (9)$$

where d_{rel} is the number of possible relations.

The bag-level and instance-level prediction tasks would be accounted for the bag loss $loss_{bag}$ and instance loss $loss_{ins}$ through categorical cross entropy. The model then calculates the final loss:

$$loss = loss_{bag} + \lambda_{ins} loss_{ins}$$

where λ_{ins} is a hyper-parameter to control the balance of instance-level loss and bag-level loss.

3.7 Soft Label

Distant supervised dataset have many noisy labels because of the assumption that each instance in a bag describes the same relation. We adopt a soft label strategy[7] to dynamically obtain a de-noised label for each bag during training process. A DS label r would be transformed to a temporary label:

$$r^{'} = \arg\max(p + \alpha \max(p) \odot L_i) \tag{10}$$

Label confidence $\alpha \in (0, 1)$ represents the reliability of DS labels. One-hot vector $L_i \in \mathbb{R}^{d_{rel}}$ indicates the DS label of a sample. $\odot$ operation represents element-wise production. p is the vector of relational scores based on the sentence or bag representation. $\max(p)$ is a hyperparameter that scaling constant which restricts the effect of the DS label. According the equation, the DS label r would be rewritten to another label only if the predicted probability of DS label is very low and the predicted probability of the most confident relation is high enough. This operation would correct partial noise data. For unintelligible samples, soft label can convert the labels into NA.

4 Experiment

4.1 Dataset

Table 1. Statistics of IPRE dataset. The dataset is splitted into training, validation and test dataset. Positive labels in the validation and the test dataset are labeled manually.

Dataset	Instance	Bag	Average Sentences
Training	287351	37948	7.57
Validation	38417	5416	7.09
Test	77092	10849	7.11

IPRE[16] is a dataset for inter-personal relationship extraction which aims to facilitate information extraction and knowledge graph construction research. In

total, IPRE has over 41,000 labeled sentences for 35 types of relations, including about 9,000 sentences annotated by workers. NA is a special relation which indicate two people have no relation. The rest of 34 relations are defined as positive relation. The dataset is generated by distant supervision, which is divided into training (70%), validation (10%) and test (20%) sets. Only the validation and test sets are labeled manually. The statistic detail of this dataset is shown in Table.1.

4.2 Metric

There are two evaluation tasks Bag-Track and Sent-Track, and both of the tracks are evaluated by the F1 score:

$$P = \frac{N_r}{N_{sys}}; R = \frac{N_r}{N_{std}}; F1 = \frac{2PR}{P+R}; \tag{11}$$

where N_r is correct relations, N_{sys} is all positive relations predicted by the system. N_{std} is all positive relations from the dataset.

4.3 Result Calculation

The model predicts a probability distribution of 35 relations $p \in \mathbb{R}^{35}$ for each bag. In general, the relation with maximum probability $l = \arg\max(p)$ should be output label. The maximum probability $c = \max(p)$ is confident probability of the label.
However, the label with low confident probability are more likely to be NA relation. We set thresholds $thres$ for all positive relations, and convert the positive label to NA label whose confidence is less than the corresponding threshold:

$$label_{bag} = \begin{cases} l & c >= thres_l \\ 0 & c < thres_l \end{cases}$$

For a positive relation rel, a relation F1 score $F1_{rel}$ could be calculated by Equ.11 with all samples which is predicted or labeled as the relation.
In order to set up reasonable threshold and improve the F1 score, we search the thresholds $thres_i$ and maximize the relation F1:

$$thres_r = \arg\max(F1_r)$$

4.4 Experiment Settings for Sent Track

We hereby give a brief description to our experiment for Sent Track, which we report here only to illustrate why jointly training with instances may work. The model setting for Sent Track largely follows the data preprocessing and feature engineering parts of the above model for Bag Track, and use a simple BERT-Base for relation classification. We also use several rule-based de-noising methods like bag smoothing after BERT model outputs the predictions.

The hyper-parameters for the best single model are: learning_rate=2.3e-5, batch_size=60, max_length=80, threshold=0.487, k=16, and we train it for 2 epoches. We experiment different hyper-parameters to obtain 13 models for ensemble.

Table 2. Result for Sent Track.

Method	Precision	Recall	F1
Bert-Base (Single)	0.366	0.436	0.398
+Rule	0.354	0.514	0.419
+Ensemble	0.402	0.488	0.441
+Ensemble with Bag	0.436	0.538	0.482

4.5 Sent Track Result

Our result for Sent Track is showed in Table 2. Our Sent Track result serves as an evidence for the effect of training on instance representation obtained from BERT: although distant supervised data is noisy, modern large pre-trained language models can still capture much more language understanding cues from sentences than traditional CNNs or LSTMs. We also report our final result for Sent Track, which is an ensemble with Bag Track prediction.

4.6 Experiment Settings for Bag Track

To avoid the overfitting by character of target name, all the name of entity pair in text are replaced by two standard name "刘伟明" (LiuWeiMing) and "李静平" (LiJingPing). BERT model has a huge number of parameters and take up a lot of GPU memory. To increase the batch size, we truncate the sentence and keep up to 55 characters. The number of instances in a bag would be limited no more than 16/32 while training/predicting. Start weights of the BERT model is [2]. The training optimizer is Adam and the learning rate is 5e-5. the dimension of the manual feature vector k is 16.
The experiment results are illustrated in Table.3. α_{ins} is the weight of instance loss. If it is set to 0, the auxilliary task would have no effect on the training and the model would be a single task model. α_{ins} and α_{bag} are the hyperparameters of soft label for instance-level task and bag-level task.

4.7 Bag Track Result

Multi-task Learning The result of model with single task proves the semantic understanding ability of BERT model. The comparison of the first two results shows that jointly training with both bag-level and instance-level improves the F1 score significantly, which is up by 3.5 percent.

Table 3. Result for Bag Track.(Single Model)

λ_{ins}	α_{ins}	α_{bag}	Precision	Recall	F1
0	0	0	0.569	0.497	0.530
1	0	0	**0.584**	**0.547**	**0.565**
1	0.85	0	0.582	0.580	0.581
1	0	0.85	0.601	0.567	0.583
1	0.85	0.85	**0.604**	**0.590**	**0.597**

Soft Label The instance-level soft label and bag-level soft label also improve the performance. Applying soft label increases F1 score by about 3 percent. Finally, both two soft labels are applied and we get the best F1 score of single model 0.597.

Ensemble Method (1)Threshold Decay: A decay factor θ is applied to reduce the thresholds in Sec.4.3 while transforming the positive label with low confidence to NA label. (2) Voting: A voting threshold δ is utilized to produce final label from the results of multiple single models. For a bag sample, if the ratio of the most positive label is greater than the voting threshold, it will be selected as the final label. Otherwise, the final label will be NA.
For our best submission record, We select 14 single models whose F1 scores are more than 0.591. We set the decay factor θ=0.75 and voting threshold δ=0.25. The F1 score of validation dataset is raised to **0.615**.

Rule-based Denoising (1)Instance Model Recheck: Recheck the positive bags (instances<=3) by the results of Sent-Track models. If all instance-level predictions in a positive bag are different from the bag-label prediction, the bag label will be replaced by NA. (2)Knowledge Inference: Many entities in training set and validation set are overlapped. We build a knowledge graph with the entity pair and relation in IPRE dataset, and implement a mechanism to inference more gender and relation information with logic rules manually designed. If any predicted result(gender or relation) conflicts with the knowledge, the label would be set to NA.
The two strategies also help to improve the final result, which increase the F1 score of validation dataset to **0.633**.

5 Conclusion and Future Work

In this paper, we present the winning model for CCKS-IPRE Bag Track, which adopts BERT as the sentence encoder and applies an auxiliary task for predicting relation of instances to improve performance. Our result reveals the potential of large language pre-trained model for distant supervised relation extraction and

shows that our proposed auxiliary task on instance level can significantly improve model performance on bag level.

On the other hand, the multi-task training scheme may also benefit model performance on instance level. One possible future work is to utilize bag-level denoising strategy to boost the performance for sentence-level relation extraction.

Acknowledgement

We thank 2019 China Conference on Knowledge Graph and Semantic Computing(CCKS2019), Suzhou University and Gowild Company for holding this contest.

References

1. Beltagy, I., Lo, K., Ammar, W.: Combining distant and direct supervision for neural relation extraction. 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). pp. 1858–1867 (2019)
2. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu, X.: Reinforcement learning for relation classification from noisy data. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
5. Jiang, X., Wang, Q., Li, P., Wang, B.: Relation extraction with multi-instance multi-label convolutional neural networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1471–1480 (2016)
6. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2124–2133 (2016)
7. Liu, T., Wang, K., Chang, B., Sui, Z.: A soft-label method for noise-tolerant distantly supervised relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1790–1795 (2017)
8. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1003–1011. Association for Computational Linguistics (2009)
9. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training
10. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8) (2019)

11. Rei, M.: Semi-supervised multitask learning for sequence labeling. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 2121–2130 (2017)
12. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 148–163. Springer (2010)
13. dos Santos, C., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 626–634 (2015)
14. Shi, P., Lin, J.: Simple bert models for relation extraction and semantic role labeling. arXiv preprint arXiv:1904.05255 (2019)
15. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 455–465. Association for Computational Linguistics (2012)
16. Wang, H., He, Z., Ma, J., Chen, W., Zhang, M.: Ipre: a dataset for inter-personal relationship extraction. arXiv preprint arXiv:1907.12801 (2019)
17. Wang, L., Cao, Z., De Melo, G., Liu, Z.: Relation classification via multi-level attention cnns (2016)
18. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
19. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237 (2019)
20. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2335–2344 (2014)
21. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)