Cost-sensitive Regularization for Label Confusion-aware Event Detection

Hongyu Lin[1,3], **Yaojie Lu**[1,3], **Xianpei Han**[1,2,*], **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
`{hongyu2016,yaojie2017,xianpei,sunle}@iscas.ac.cn`

Abstract

In supervised event detection, most of the mislabeling occurs between a small number of confusing type pairs, including trigger-NIL pairs and sibling sub-types of the same coarse type. To address this label confusion problem, this paper proposes cost-sensitive regularization, which can force the training procedure to concentrate more on optimizing confusing type pairs. Specifically, we introduce a cost-weighted term into the training loss, which penalizes more on mislabeling between confusing label pairs. Furthermore, we also propose two estimators which can effectively measure such label confusion based on instance-level or population-level statistics. Experiments on TAC-KBP 2017 datasets demonstrate that the proposed method can significantly improve the performances of different models in both English and Chinese event detection.

1 Introduction

Automatic event extraction is a fundamental task in information extraction. Event detection, aiming to identify trigger words of specific types of events, is a vital step of event extraction. For example, from sentence "Mary was injured, and then she died", an event detection system is required to detect a *Life:Injure* event triggered by "injured" and a *Life:Die* event triggered by "died".

Recently, neural network-based supervised models have achieved promising progress in event detection (Nguyen and Grishman, 2015; Chen et al., 2015; Ghaeini et al., 2016). Commonly, these methods regard event detection as a word-wise classification task with one *NIL* class for tokens do not trigger any event. Specifically, a neural network automatically extracts high-level features and then feed them into a classifier to categorize words into their corresponding event sub-types (or *NIL*). Optimization criteria of such models often involves in minimizing cross-entropy loss, which equals to maximize the likelihood of making correct predictions on the training data.

	BC	CT	CR	MT	NIL	CC
BC	41.3	14.4	2.1	1.6	39.0	1.7
CT	8.5	42.7	4.7	2.6	40.6	0.9
CR	5.7	7.3	50.0	1.1	32.3	2.9
MT	3.0	7.7	6.1	28.7	51.3	3.2

Table 1: Prediction percentage heatmap of triggers with *Contact* coarse type. Row labels are the golden label and the column labels indicate the prediction. BC: Broadcast; CT: Conctact(sub-type); CR: Correspondence; MT: Meet; CC: Other cross coarse-type errors.

However, we find that in supervised event detection, most of the mislabeling occurs between a small number of confusing type pairs. We refer to this phenomenon as *label confusion*. Specifically, there are mainly two types of label confusion in event detection: 1) trigger/NIL confusion; 2) sibling sub-types confusion. For example, both *Transaction:Transfer-money* and *Transaction:Transfer-ownership* events are frequently triggered by word "give". Besides, in many cases "give" does not serve as a trigger word. Table 1 shows the classification results of a state-of-the-art event detection model (Chen et al., 2015) on all event triggers with coarse type of *Contact* on TAC-KBP 2017 English Event Detection dataset. We can see that the model severely suffers from two types of label confusion mentioned above: more than 50% mislabeling happens between trigger/NIL decision due to the ambiguity of natural language. Furthermore, the majority of remaining errors are between sibling sub-types of the same coarse type because of their semantic relatedness (Liu et al., 2017b). Similar results are also observed in other event detection datasets such as ACE2005 (Liu et al., 2018a). Therefore,

[*] Corresponding author.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5278–5283
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

it is critical to enhance the supervised event detection models by taking such label confusion problem into consideration.

In this paper, inspired by cost-sensitive learning (Ling and Sheng, 2011), we introduce cost-sensitive regularization to model and exploit the label confusion during model optimization, which can make the training procedure more sensitive to confusing type pairs. Specifically, the proposed regularizer reshapes the loss function of model training by penalizing the likelihood of making wrong predictions with a cost-weighted term. If instances of class i are more frequently misclassified into class j, we assign a higher cost to this type pair to make the model intensively learn to distinguish between them. Consequently, the training procedure of models not only considers the probability of making correct prediction, but also tries to separate confusing type pairs with a larger margin. Furthermore, in order to estimate such cost automatically, this paper proposes two estimators based on population-level or instance-level statistics.

We conducted experiments on TAC-KBP 2017 Event Nugget Detection datasets. Experiments show that our method can significantly reduce the errors between confusing type pairs, and therefore leads to better performance of different models in both English and Chinese event detection. To the best of our knowledge, this is the first work which tackles with the label confusion problem of event detection and tries to address it in a cost-sensitive regularization paradigm.

2 Cost-sensitive Regularization for Neural Event Detection

2.1 Neural Network Based Event Detection

The state-of-the-art neural network models commonly transform event detection into a word-wise classification task. Formally, let $D = \{(x_i, y_i)|i = 1, 2, ..., n\}$ denote n training instances, $P(y|x;\theta)$ is the neural network model parameterized by θ, which takes representation (feature) x as input and outputs the probability that x is a trigger of event sub-type y (or *NIL*). Training procedure of such models commonly involves in minimizing following cross-entropy loss:

$$\mathcal{L}_{CE}(\theta) = -\sum_{(x_i,y_i)\in D} \log P(y_i|x_i;\theta) \tag{1}$$

which corresponds to maximize the log-likelihood of the model making the correct prediction on all training instances and does not take the confusion between different type pairs into consideration.

2.2 Cost-sensitive Regularization

As discussed above, the key to improve event detection performance is to solve the label confusion problem, i.e., to guide the training procedure to concentrate on distinguishing between more confusing type pairs such as trigger/NIL pairs and sibling sub-event pairs. To this end, we propose cost-sensitive regularization, which reshapes the training loss with a cost-weighted term of the log-likelihood of making wrong prediction. Formally, the proposed regularizer is defined as:

$$\mathcal{L}_{CS}(\theta) = \sum_{(x_i,y_i)\in D} \sum_{y_j \neq y_i} C(y_i, y_j; x_i) \log P(y_j|x_i;\theta) \tag{2}$$

where $C(y_i, y_j; x)$ is a positive cost of mislabeling an instance x with golden label y_i into label y_j. A higher $C(y_i, y_j; x)$ is assigned if y_i and y_j is a more confusing type pair (i.e., more easily mislabeled by the current model). Therefore, the cost-sensitive regularizer will make the training procedure pay more attention to distinguish between confusing type pairs because they have larger impact on the training loss. Finally, the entire optimization objective can be written as:

$$\mathcal{L}(\theta) = \mathcal{L}_{CE}(\theta) + \lambda \mathcal{L}_{CS}(\theta) \tag{3}$$

where λ is a hyper-parameter that controls the relative impact of our cost-sensitive regularizer.

3 Cost Estimation

Obviously it is critical for the proposed cost-sensitive regularization to have an accurate estimation of the cost $C(y_i, y_j; x)$. In this section, we propose two approaches for this issue based on population-level or instance-level statistics.

3.1 Population-level Estimator

A straightforward approach for measuring such costs is to use the relative mislabeling risk on the dataset. Therefore our population-level cost estimator is defined as:

$$C_{POP}(y_i, y_j; x_i) = \frac{\#(y_i, y_j)}{\sum_j \#(y_i, y_j)} \tag{4}$$

where $\#(y_i, y_j)$ is the number of instances with golden label y_i but being classified into class y_j in the corpus. These statistics can be computed either on the training set or on the development set. This paper uses statistics on development set due to its compact size. And the estimators are updated every epoch during the training procedure.

3.2 Instance-level Estimator

The population-level estimators requires large computation cost to predict on the entire dataset when updating the estimators. To handle this issue, we propose another estimation method based directly on instance-level statistics. Inspire by Lin et al. (2017), the probability $P(y_j|x_i;\theta)$ of classifying instance x_i into the wrong class y_j can be directly regarded as the mislabeling risk of that instance. Therefore our instance-level estimator is:

$$C_{INS}(y_i, y_j; x_i) = P(y_j|x_i;\theta) \tag{5}$$

Then cost-sensitive regularizer for each training instance can be written as:

$$\mathcal{L}_{INS}(x_i;\theta) = \sum_{y_j \neq y_i} P(y_j|x_i;\theta) \log P(y_j|x_i;\theta) \tag{6}$$

Note that if the probability of making correct prediction (i.e., $P(y_i|x_i;\theta)$) is fixed, $\mathcal{L}_{INS}(x_i;\theta)$ achieves its minimum when the probabilities of mislabeling x_i into all incorrect classes are equal. This is equivalent to maximize the margin between the probability of golden label and that of any other class. In this circumstance, the loss $\mathcal{L}(\theta)$ can be regarded as a combination of maximizing both the likelihood of correct prediction and the margin between correct and incorrect classes.

4 Experiments

4.1 Experimental Settings

We conducted experiments on both English and Chinese on TAC-KBP 2017 Event Nugget Detection Evaluation datasets (LDC2017E55). For English, previously released RichERE corpus, including LDC2015E29, LDC2015E68, LDC2016E31 and the English part of LDC2017E02, were used for training. For Chinese, LDC2015E105, LDC2015E112, LDC2015E78 and the Chinese part of LDC2017E02 were used. For both English and Chinese, we sampled 20 documents from LDC2017E02 as the development set. Finally, there were 866/20/167 documents and 506/20/167 documents in English and Chinese train/development/test set respectively.

We conducted experiments on two state-of-the-art neural network event detection models to verify the portability of our method. One is DMCNN model proposed by Chen et al. (2015). Another is a LSTM model by Yang and Mitchell (2017). Due to page limitation, please refer to original papers for details.

Model	English			Chinese		
	P	R	F1	P	R	F1
			LSTM			
CE	73.46	34.23	46.70	70.35	35.43	47.13
Focal	69.20	38.71	49.64	68.10	35.76	46.90
Hinge	62.51	44.36	51.89	58.34	43.40	49.77
Sampling	58.57	48.26	52.92	57.61	44.54	50.24
CR-POP	62.35	46.98	53.58	53.18	49.55	51.30
CR-INS	58.64	49.55	**53.71**	49.19	55.83	**52.30**
			DMCNN			
CE	75.15	34.16	47.00	73.50	35.81	48.16
Focal	70.68	37.63	49.11	69.04	38.87	49.74
Hinge	67.49	42.67	52.28	60.27	45.50	51.85
Sampling	64.05	45.08	52.91	54.85	50.35	52.50
CR-POP	64.82	45.73	53.63	55.89	50.81	53.23
CR-INS	64.74	46.14	**53.88**	54.91	51.93	**53.38**

Table 2: Overall results. *CR-POP* and *CR-INS* are our method with population-level and instance-level estimators. All F1 improvements made by *CR-POP* and *CR-INS* are statistically significant with $p < 0.05$.

4.2 Baselines[1]

Following baselines were compared:

1) **Cross-entropy Loss (CE)**, the vanilla loss.

2) **Focal Loss (Focal)** (Lin et al., 2017), which is an instance-level method that rescales the loss with a factor proportional to the mislabeling probability to enhance the learning on hard instances.

3) **Hinge Loss (Hinge)**, which tries to separate the correct and incorrect predictions with a margin larger than a constant and is widely used in many machine learning tasks.

4) **Under-sampling (Sampling)**, a representative cost-sensitive learning approaches which samples instances balance the model learning and is widely used in event detection to deal with imbalance (Chen et al., 2015).

We also compared our methods with the top systems in TAC-KBP 2017 Evaluation. We evaluated all systems with micro-averaged Precision(P), Recall(R) and F1 using the official toolkit[2].

4.3 Overall Results

Table 2 shows the overall performance on TAC-KBP 2017 datasets. We can see that:

1) **Cost-sensitive regularization can significantly improve the event detection performance by taking mislabeling costs into consideration.** The proposed CR-INS and the CR-POP

[1] Our source code and hyper-parameter configures are openly available at `github.com/sanmusunrise/CSR`.

[2] `github.com/hunterhector/EvmEval`

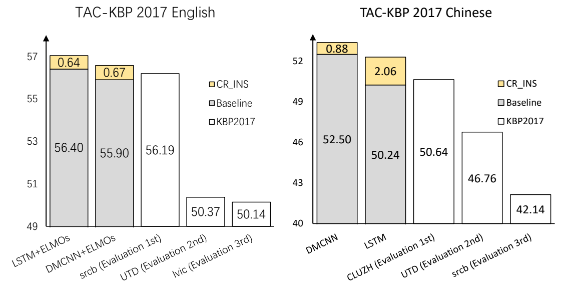

Figure 1: Comparison with the top systems in TAC-KBP 2017. CR is our CR-INS method. The *srcb* system in English used additional CRF based models to deal with multi-word triggers in English, which is not considered in our model and leads to a significant higher recall than other competitors.

steadily outperform corresponding baselines. Besides, compared with population-level estimators, instance-level cost estimators are more effective. This may because instance-level estimators can be updated every batch while population-level estimators are updated every epoch, which leads to a more accurate estimation.

2) **Cost-sensitive regularization is robust to different languages and models.** We can see that cost-sensitive regularization achieves significant improvements on both English and Chinese datasets with both CNN and RNN models. This indicates that our method is robust and can be applied to different models and datasets.

3) **Data imbalance is not the only reason behind label confusion.** Even Focal and Sampling baselines deals with the data imbalance problem, they still cannot achieve comparable performance with CR-POP and CR-INS. This means that there are still other reasons which are not fully resolved by conventional methods for data imbalance.

4.4 Comparing with State-of-the-art Systems

Figure 1 compares our models with the top systems in TAC-KBP 2017 Evaluation. To achieve a strong baseline[3], we also incorporate ELMOs (Peters et al., 2018) to English system for better representations. We can see that CR-INS can further gain significant improvements over all strong baselines which have already achieved comparable performance with top systems. In both English and Chinese, CR-INS achieves the new SOTA performance, which demonstrates its effectiveness.

Error Rate (%)	**SP**	**CR**	Δ
Total Error	42.97	38.84	-9.6%
- Trigger/NIL	33.39	31.15	-6.7%
- Sibling Sub-types	8.15	6.25	-23.3%
- Other	1.43	1.44	+0.6%

Table 3: Error rates (CNN) on trigger words on the Chinese test set with Sampling(SP) and CR-INS(CR).

4.5 Error Analysis

To clearly show where the improvement of our method comes from, we compared the mislabeling made by Sampling and our CR-INS method. Table 3 shows the results. We can first see that trigger/NIL mislabeling and sibling sub-types mislabeling make up most of errors of CE baseline. This further verifies our motivation. Besides, cost-sensitive regularization significantly reduces these two kinds of errors without introducing more other types of mislabeling, which clearly demonstrates the effectiveness of our method.

5 Related Work

Neural Network based Event Detection. Recently, neural network based methods have achieved promising progress in event detection, especially with CNNs (Chen et al., 2015; Nguyen and Grishman, 2015) and Bi-LSTMs (Zeng et al., 2016; Yang and Mitchell, 2017) based models as automatic feature extractors. Improvements have been made by incorporating arguments knowledge (Nguyen et al., 2016; Liu et al., 2017a; Nguyen and Grishman, 2018; Hong et al., 2018) or capturing larger scale of contexts with more complicated architectures (Feng et al., 2016; Nguyen and Grishman, 2016; Ghaeini et al., 2016; Lin et al., 2018a,b; Liu et al., 2018a,b; Sha et al., 2018; Chen et al., 2018).

Cost-sensitive Learning. Cost-sensitive learning has long been studied in machine learning (Elkan, 2001; Zhou, 2011; Ling and Sheng, 2011). It can be applied both at algorithm-level (Anand et al., 1993; Domingos, 1999; Sun et al., 2007; Krawczyk et al., 2014; Kusner et al., 2014) or data-level (Ting, 2002; Zadrozny et al., 2003; Mirza et al., 2013), which has achieved great success especially in learning with imbalanced data.

6 Conclusions

In this paper, we propose cost-sensitive regularization for neural event detection, which introduces a cost-weighted term of mislabeling likelihood

[3] Top systems in the evaluation are commonly ensembling models with additional resources, while reported in-house results are of single model.

to enhance the training procedure to concentrate more on confusing type pairs. Experiments show that our methods significantly improve the performance of neural network event detection models.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. Moreover, this work is supported by the National Natural Science Foundation of China under Grants no. 61433015, 61572477 and 61772505; the Projects of the Chinese Language Committee under Grants no. WT135-24; and the Young Elite Scientists Sponsorship Program no. YESS20160177.

References

Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. 1993. An improved algorithm for neural network classification of imbalanced training sets. *IEEE Transactions on Neural Networks*, 4(6):962–969.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL 2015*.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multilevel attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1267–1276. Association for Computational Linguistics.

Pedro M. Domingos. 1999. Metacost: A general method for making classifiers cost-sensitive. In *KDD*.

Charles Elkan. 2001. The foundations of cost-sensitive learning. In *IJCAI 2001*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Bing Qin, Heng Ji, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of ACL 2016*.

Reza Ghaeini, Xiaoli Z Fern, Liang Huang, and Prasad Tadepalli. 2016. Event nugget detection with forward-backward recurrent neural networks. In *Proceedings of ACL 2016*.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Qiaoming Zhu, and Guodong Zhou. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 515–526. Association for Computational Linguistics.

Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562.

Matt J Kusner, Wenlin Chen, Quan Zhou, Zhixiang Eddie Xu, Kilian Q Weinberger, and Yixin Chen. 2014. Feature-cost sensitive learning with submodular trees of classifiers. In *AAAI 2014*.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2018a. Adaptive scaling for sparse detection in information extraction. *arXiv preprint arXiv:1805.00250*.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2018b. Nugget proposal networks for chinese event detection. *arXiv preprint arXiv:1805.00249*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Charles X Ling and Victor S Sheng. 2011. Cost-sensitive learning. In *Encyclopedia of machine learning*, pages 231–235. Springer.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018a. Event detection via gated multilingual attention mechanism. In *Proceedings of AAAI2018*.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017a. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of ACL2017*.

Shulin Liu, Yubo Chen, Kang Liu, Jun Zhao, Zhunchen Luo, and Wei Luo. 2017b. Improving event detection via information sharing among related event types. In *CCL 2017*, pages 122–134.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018b. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256. Association for Computational Linguistics.

Bilal Mirza, Zhiping Lin, and Kar-Ann Toh. 2013. Weighted online sequential extreme learning machine for class imbalance learning. *Neural processing letters*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL-HLT 2016*.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of ACL 2015*.

Thien Huu Nguyen and Ralph Grishman. 2016. Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of EMNLP 2016*.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of AAAI2018*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of AAAI2018*.

Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378.

Kai Ming Ting. 2002. An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):659–665.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *Proceedings of ACL2017*.

Bianca Zadrozny, John Langford, and Naoki Abe. 2003. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM 2003*, pages 435–442.

Ying Zeng, Honghui Yang, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2016. A convolution bilstm neural network model for chinese event extraction. In *Proceedings of NLPCC-ICCPOL 2016*.

Zhi-Hua Zhou. 2011. Cost-sensitive learning. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 17–18. Springer.