

SELCOR: SELF-CORRECTION FOR WEAKLY SUPERVISED LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Powerful machine learning models often require training with large amounts of labeled data. Collecting precise labels from human supervision is expensive and time-consuming. Instead, it is much easier to obtain large-scale weak labels from multiple weak supervision sources such as rules or knowledge graphs, whereas weak labels could be noisy and make models prone to overfitting. We propose a self-training method for weakly supervised learning without using any true label. Our method learns a joint model with a corrector and a predictor, where the predictor generates pseudo clean labels and the corrector revises weak labels to *reproduces* the pseudo labels generated by the predictor. The joint model is trained by encouraging consistency between label generation and label correction, such that the predictor and corrector can iteratively improve each other to generate more reliable pseudo label for self-training. In this way, our method makes full use of weak labels and effectively suppresses label noise in weak labels and pseudo labels. Experiments on 8 benchmark datasets show that our method outperforms existing weakly supervised methods by large margins.

1 INTRODUCTION

The great success of state-of-the-art machine learning models such as deep neural networks relies on large-scale labeled data for training. Obtaining labeled data with human annotation is typically expensive and time-consuming, which becomes a major bottleneck in deploying machine learning models in real-world applications. Weak supervision that collects weak labels from multiple noisy sources such as knowledge graphs Hoffmann et al. (2011); Liang et al. (2020), crowd-sourcing Karger et al. (2011); Dalvi et al. (2013), and domain-specific rules Hu et al. (2016); Safranchik et al. (2020), has emerged as an efficient way of data annotation, and shown to be useful in a diverse set of tasks including image classification, relation extraction, and named entity recognition, to name a few.

Due to the heuristic and noisy nature of weak labels, learning from weak supervision often suffers from label noise, which can make *instance-dependent* errors and mislead the model training. The main task of weakly supervised learning is to discover the unobserved true labels by leveraging weak labels in a noise-robust way. Previous work achieves this by multi-source label aggregation Ratner et al. (2019a;b) and noise transition learning Luo et al. (2017); Wang et al. (2019). Recently, self-training has been *adopted* for weakly supervised learning Hu et al. (2019); Ren et al. (2020); Karamanolakis et al. (2021); Yu et al. (2021). These self-training based methods start learning from weak labels and then iteratively generate and retrain with pseudo (clean) labels for denoising, which leads to the state-of-the-art performance of weakly supervised learning.

Despite its success, self-training is sensitive to label noise and tends to accumulate errors from wrong labels Guo et al. (2017), leading to gradually deteriorated performance as the number of iterations increases. To address this problem, existing methods commonly suppress label noise by filtering or down-weighting data with possibly wrong or low-confident labels Shu et al. (2019); Ren et al. (2020); Yu et al. (2021). However, denoising via removing or down-weighting require careful tuning a threshold or sample weights for good performance and may cause loss of information, which is undesirable for weak supervision. Some other approaches Hendrycks et al. (2018); Zheng et al. (2021) aim to correct weak labels by learning a denoising module, whereas they have to know extra side information or ground-truth labels in advance for training the denoising module.

In this paper, we propose SELCOR, a self-training based method for weakly supervised learning without using any ground-truth label. Our method is motivated by the hypothesis that a good weakly supervised model should not only learn *what is the correct label* but also *how wrong labels should be corrected*. However, when the ground-truth labels are not available, both of these goals are non-trivial to achieve. Instead of directly estimating the true labels, we propose to iteratively generate and update pseudo labels by encouraging *consistency* between the label generation and label correction. Specifically, SELCOR jointly learns a predictor and a corrector upon a shared encoder, where the predictor generates pseudo clean labels and the corrector revises weak labels to *reproduces* the pseudo labels generated by the predictor. The corrector is instantiated by a principled noise model that performs label correction by estimating the probabilities of weak labels to be false positive and false negative. Although the pseudo labels could change during self-training, there should be a consensus between the predictor and the corrector if the pseudo label is correct for a given data sample, which leads to a suitable objective for weakly supervised learning. By exploiting the consistency between label generation and label correction, this self-correction mechanism can effectively suppress error accumulation and facilitate self-training with weak supervision. In our method, the corrector is used to capture the correlations between the pseudo labels and the weak labels, which helps the predictor to generate more reliable pseudo labels rather than fixing errors in weak labels. In addition, our method makes full use of the training set without removing or reweighting any training samples. These are quite different from existing approaches that try to *avoid using low-confident weak labels* for training or utilize ground-truth labels to correct wrong weak labels.

Our main contributions include: 1) We propose the self-correction paradigm to learn from weak labels via encouraging consistency between label generation and label correction without using any ground-truth label. 2) We propose a noise model for label correction by estimating the probabilities of weak labels to be false positive and false negative. 3) We compare our method on 8 benchmark datasets and show that our method greatly benefits from the proposed self-correction paradigm and outperforms existing weakly supervised methods by large margins.

2 SELCOR: WEAK SUPERVISION VIA SELF-CORRECTION

In this section, we provide the background of weak supervision and present our SELCOR method for weakly supervised learning.

2.1 PROBLEM SETUP

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be the set of N training samples, where $y_i \in \mathcal{Y} = \{1, \dots, C\}$ are the *unknown* labels for C classes. In our setting, the true labels y_i are unknown, and we only have *weak labels* $\tilde{y}_i \in \tilde{\mathcal{Y}} = \{0, 1, \dots, C\}$ obtained from weak supervision sources such as rules or knowledge graphs, where $\tilde{y}_i = 0$ means that all the weak supervision sources *abstain* from labeling \mathbf{x}_i . For training samples that have multiple weak labels, we simply apply majority voting to aggregate different weak supervision information into one weak label. Since weak labels inevitably contain errors and noise, directly learning from weak labels usually leads to degraded performance. Our goal is to learn a noise-aware model from weak labels that can perform well on the unseen test set.

2.2 SELF-CORRECTION FRAMEWORK

Label generation with the predictor. We start with training a predictor $f = p(y_i|\mathbf{x}_i) = (h_{\text{pred}} \circ \psi)(\mathbf{x}_i)$, where ψ is the encoder that generates latent representations of training samples. h_{pred} is the prediction head that projects $\psi(\mathbf{x}_i)$ into a C -dimensional label space, where each dimension represents the prediction confidence for the corresponding class. The predictor f models the probability $p(y_i|\mathbf{x}_i)$ and can be trained on weak labels by minimizing the cross-entropy loss as follows:

$$L_{\text{init}}(h_{\text{pred}}, \psi) = \frac{1}{N} \sum_{i=1}^N \text{CE}((h_{\text{pred}} \circ \psi)(\mathbf{x}_i), \tilde{y}_i). \quad (1)$$

Since weak labels are generally noisy, training from weak labels tends to memorize wrongly given labels Zhang et al. (2021a), leading to poor generalization performance. On the other hand, it has been observed that deep networks would first memorize simple patterns (e.g., training data of clean

labels) and then noise data Arpit et al. (2017). Motivated by the memorization effects of deep neural networks, a model that learns data patterns before memorizing noisy labels can serve as a supervision source and assign pseudo labels to each training sample. Those pseudo labels can be exploited together with weak labels to improve the performance of weakly supervised learning. However, self-training itself suffers from label noise in the pseudo labels, leading to deteriorated performance. To reduce the pseudo label noise, existing approaches denoise pseudo labels by filtering or down-weighting data with possibly wrong or low-confident labels Yu et al. (2021).

Label correction with the corrector. Unlike previous self-training based approaches that focus only on finding high quality pseudo labels, we propose a new self-training paradigm for weakly supervised learning called SELf-CORrection (SELCOR), which focuses on learning *the correlations between label generation and label correction*. We hypothesize that a good weakly supervised model should not only predict (generate) the correct labels but also be able to correct noisy weak labels. To achieve this, we introduce a correction head upon the encoder ψ that models the correcting posterior probability $p(y_i|\tilde{y}_i, \mathbf{x}_i)$. Intuitively, $p(y_i|\tilde{y}_i, \mathbf{x}_i)$ indicates what is the true label of input sample \mathbf{x}_i when observing a weak label \tilde{y}_i . Ideally, the corrector should be trained on the true labels, whereas this is inapplicable in the pure weakly supervised learning setting. Although the true labels are unknown, the corrector and the predictor are still closely related in the sense that they should produce consistent outputs for the same input \mathbf{x}_i . Therefore, we let the corrector reproduce the pseudo labels generated by the predictor, and jointly update the predictor $p(y_i|\mathbf{x}_i)$ and the corrector $p(y_i|\tilde{y}_i, \mathbf{x}_i)$ by encouraging their consistency. This captures the correlations between weak labels and pseudo labels, and enables the predictor and the corrector improve each other iteratively, which in turn leads to more reliable pseudo labels for self-training. The proposed SELCOR model is illustrated in Figure 1.

2.3 CORRECTING POSTERIOR MODELING

Now the key question is how to model the corrector and the correcting posterior distribution. We achieve this by proposing a principled noise model under the probabilistic framework. Specifically, we reduce multi-class classification into one binary classification problem per class. In this way, there are only two types of errors in weak labels for each class, i.e., *false positive* and *false negative* errors.

Let $\mathbf{y}_i, \tilde{\mathbf{y}}_i \in \mathbb{R}^C$ be the one-hot label vectors of the true label y_i and the weak label \tilde{y}_i , respectively, and $y_{ji} \in \{0, 1\}$ be the binary label of the j -th class from \mathbf{y}_i . We can approximate the probability of weak label \tilde{y}_{ji} to be false positive and false negative by $p(y_{ji} = 0|\tilde{y}_{ji} = 1, \mathbf{x}_i) = (h_{fp} \circ \psi)(\mathbf{x}_i)$ and $p(y_{ji} = 1|\tilde{y}_{ji} = 0, \mathbf{x}_i) = (h_{fn} \circ \psi)(\mathbf{x}_i)$, respectively, where ψ is the encoder shared with the predictor, h_{fp} and h_{fn} are the projection functions, and $(h_{fp} \circ \psi)_j(\mathbf{x}_i)$ represents the j -th component of the output vector. Then, the correcting posterior for each class can be formulated as follows:

$$p(y_{ji}|\tilde{y}_{ji}, \mathbf{x}_i) = (1 - p(y_{ji} = 0|\tilde{y}_{ji} = 1, \mathbf{x}_i)) \cdot \tilde{y}_{ji} + p(y_{ji} = 1|\tilde{y}_{ji} = 0, \mathbf{x}_i) \cdot (1 - \tilde{y}_{ji}). \quad (2)$$

Essentially, $p(y_{ji}|\tilde{y}_{ji}, \mathbf{x}_i)$ is an instance-dependent noise model derived from the probability calculation $p(y|\mathbf{x}) = \int p(y|\tilde{y}, \mathbf{x})p(\tilde{y})d\tilde{y}$ in the binary case, which transforms \tilde{y}_{ji} to its true label given \mathbf{x}_i . When it comes to self-correction, learning $p(y_{ji}|\tilde{y}_{ji}, \mathbf{x}_i)$ helps the predictor to generate pseudo labels that are consistent with the rectified weak labels via (2).

Weak label smoothing. For a typical multi-class classification problem, most weak labels $\tilde{y}_{ji} = 0$. In this case, the noise model (2) could have the underfitting problem since $p(y_{ji} = 0|\tilde{y}_{ji} = 1, \mathbf{x}_i)$ will always have zero gradients when $\tilde{y}_{ji} = 0$. To address this problem, we propose weak label smoothing as follows:

$$\tilde{y}_{ji}^s = \tilde{y}_{ji} \cdot (1 - \alpha) + \alpha/C, \quad (3)$$

where α is the smoothing parameter. Weak label smoothing essentially injects small constant noise into weak labels and transforms hard weak labels into soft ones. Such soft weak labels facilitate gradient updates for learning the noise model, and allow all the false-positive probabilities $p(y_{ji} = 0|\tilde{y}_{ji} = 1, \mathbf{x}_i)$ to be updated when $\tilde{y}_{ji} = 0$. Please note that weak label smoothing is only used for constructing the noise model (2). This is different from the classical label smoothing strategy that modifies targets of loss functions for improving the model generalization.

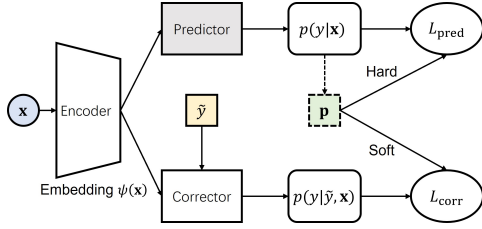


Figure 1: The proposed SELCOR model. For a sample \mathbf{x} with weak label \tilde{y} , SELCOR generates pseudo label \mathbf{p} for \mathbf{x} from the previous model checkpoint, and learns the distributions $p(y|\mathbf{x})$ and $p(y|\tilde{y}, \mathbf{x})$ by jointly minimizing L_{pred} and L_{corr} .

Algorithm 1 Training Procedure for SELCOR.

Input: Training set $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^N$.
for sampled minibatch \mathcal{B} from \mathcal{D} **do**
 # Initialize the model with weak labels.
 Update h_{pred} and ψ to minimize (1).
end for
for $t = 1, 2, \dots, T$ **do**
 # Update pseudo labels for every T' steps.
 if $t \bmod T' == 0$ **then**
 Update pseudo labels \mathbf{p}_i via (4).
 end if
 Sample a minibatch \mathcal{B} from \mathcal{D} .
 # Joint classification and label correction.
 Update h_{fp} , h_{fn} and ψ via (7).
end for
Output: Learned SELCOR model.

2.4 MODEL TRAINING

In this section, we present how to train our SELCOR model along with several tricks for better performance. We first initialize the predictor f by minimizing the standard cross-entropy loss (1) with weak labels. At this step, the corrector will not be learned. Then the initialized predictor can be used to generate pseudo labels for each training sample \mathbf{x}_i .

Pseudo label generation. Instead of directly using the outputs of f as pseudo labels, we employ the batch-wise normalized pseudo labels $\mathbf{p}_i \in \mathbb{R}^C$ Xie et al. (2016) for self-training as follows:

$$p_{ji} = \frac{f_j^2(\mathbf{x}_i)/Z_j}{\sum_{j'=1}^C f_{j'}^2(\mathbf{x}_i)/Z_{j'}}, Z_j = \sum_{\mathbf{x}' \in \mathcal{B}} f_j^2(\mathbf{x}'), \quad (4)$$

where p_{ji} is the pseudo label of the j -th class for \mathbf{x}_i , $f_j^2(\mathbf{x}_i)$ is the square of the j -th component of $f(\mathbf{x}_i) \in \mathbb{R}^C$, and Z_j is the sum of $f_j^2(\mathbf{x}')$ over batch \mathcal{B} . This normalization strategy works well in our experiments, and has been used by other self-training methods Yu et al. (2021). We also test using the corrector to generate pseudo labels for self-training but find that the model tends to overfit the low-quality pseudo labels generated at the first few iterations. This motivates us to separate the corrector from label generation, leading to good estimations of the noise model and thus better performance in our experiments.

Joint classification and label correction. We update the model parameters with the pseudo labels p_{ji} by jointly learning the distributions $p(y_i|\tilde{y}_i, \mathbf{x}_i)$ and $p(y_i|\mathbf{x}_i)$ for self-correction and prediction, respectively. For learning corrector $p(y_i|\tilde{y}_i, \mathbf{x}_i)$, we use the binary cross-entropy loss as follows:

$$L_{\text{corr}}(h_{\text{fp}}, h_{\text{fn}}, \psi) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C \text{BCE}(p(y_{ji}|\tilde{y}_{ji}^s, \mathbf{x}_i), p_{ji}), \quad (5)$$

where $p(y_{ji}|\tilde{y}_{ji}^s, \mathbf{x}_i) = (1 - (h_{\text{fp}} \circ \psi)(\mathbf{x}_i)) \cdot \tilde{y}_{ji}^s + (h_{\text{fn}} \circ \psi)(\mathbf{x}_i) \cdot (1 - \tilde{y}_{ji}^s)$ is obtained from the noise model (2). For learning predictor $p(y_i|\mathbf{x}_i)$, we use the following cross-entropy loss

$$L_{\text{pred}}(\psi) = \frac{1}{N} \sum_{i=1}^N \text{CE}((h_{\text{pred}} \circ \psi)(\mathbf{x}_i), \hat{p}_i), \quad (6)$$

where $\hat{p}_i = \arg \max \mathbf{p}_i$ is the hard version of the pseudo label \mathbf{p}_i . When the model is not well trained at early iteration, the predictor f tend to generate less discriminative pseudo labels with low confidence. Training on such less informative pseudo labels makes the model get stuck into bad local solutions. To address this problem, we use the *hard* pseudo label \hat{p}_i for learning the predictor. This trick encourages the model to generate more confident pseudo labels and achieves much better performance in our experiments.

Fixed predictor. Recall that the goal of the corrector is to transform weak label \tilde{y} into the pseudo label $p = (h_{pred} \circ \phi)(\mathbf{x})$ via the noise model (2). In other words, we hope the corrector $(h_{corr} \circ \phi)(\mathbf{x}, \tilde{y}) \approx p = (h_{pred} \circ \phi)(\mathbf{x})$. However, when both the predictor and corrector are free for updating, the optimal solutions are not unique, as there could exist new sets of prediction and correction heads that also satisfy this equation. To make the correlations between the weak label \tilde{y}_{ji} and the pseudo label more identifiable, we propose to simplify the estimation of the corrector by fixing the predictor h_{pred} after initialization. This simple modification leads to consistent improvements. Some recent works Chen et al. (2022); Yang et al. (2022) also show that fixing the prediction head can reduce error propagation in self-training. We study the empirical effect of fixing h_{pred} in Sec. 4.5.

Putting (5) and (6) together, the training objective of SELCOR is:

$$\min_{h_{fp}, h_{fn}, \psi} L_{corr}(h_{fp}, h_{fn}, \psi) + \gamma L_{pred}(\psi), \quad (7)$$

where γ is the trade-off parameter. Finally, we perform self-training by updating the pseudo-labels and the model parameters iteratively. For testing, the prediction of input \mathbf{x} is obtained from the predictor $f = (h_{pred} \circ \psi)(\mathbf{x})$. We provide the pseudo code for training SELCOR in Algorithm 1.

3 RELATED WORK

Multi-source weak supervision. Learning from noisy labels has been extensively studied in the machine learning community Khetan et al. (2018); Zhang & Sabuncu (2018); Ren et al. (2018). One representative method is data programming Ratner et al. (2016) which aims to aggregate noisy signals from multiple supervision sources. Along this line, many attempts have been made to develop a label model to generate denoised weak labels from weak supervision sources Ratner et al. (2019a;b). Label models such as majority voting and Snorkel Ratner et al. (2019a) fail to consider the end model and downstream tasks in training. Further improvements have been proposed to jointly train a label model and a downstream model by estimating the reliability of label sources Ren et al. (2020); Awasthi et al. (2020); Karamanolakis et al. (2021); Rühling Cachay et al. (2021); Mazzetto et al. (2021); Arachie & Huang (2021; 2022). Our method is orthogonal to these data programming methods, and is flexible to take weak labels from any label models for self-training.

Noise transition learning. Another prominent way of learning from noisy labels is to estimate a noise transition matrix \mathbf{T} that capture relationships between clean and noisy labels, where \mathbf{T}_{ij} represent the probability $p(\tilde{y} = j | y = i, \mathbf{x})$ for the weak label \tilde{y} of class j given the true label y is class i . Such transition matrix can be used to re-label training samples Cheng et al. (2020) or adapt a noise-aware loss function Patrini et al. (2017) for improved performance. However, estimating \mathbf{T} usually requires assuming that the transition is instance-independent, i.e., $p(\tilde{y} = j | y = i, \mathbf{x}) = p(\tilde{y} = j | y = i)$ or making use of clean labels, which are inapplicable in our weak supervision setting. Our method is closely related to these noise transition-based methods Luo et al. (2017); Wang et al. (2019) in the sense that we all construct a noise model to learn the transition from weak labels. The main differences of these methods from ours are: 1) Our method aims to learn the correlations between weak labels and dynamically changed pseudo labels rather than fixed true labels, which is a much more difficult problem and requires non-trivial designs and implementations. 2) The corrector in our model is used as an auxiliary task rather than directly generating corrected labels. 3) Existing methods rely on clean labels to learn the transition matrix \mathbf{T} , while our method does not need to use any clean label.

Self-training. Self-training seeks to use model predictions as pseudo labels to update the model itself gradually and has become a classic technique for learning from low-resource supervision Yarowsky (1995); Nigam & Ghani (2000); Lee (2013). Recently, self-training has been successfully applied for weakly supervised learning Hu et al. (2019); Ren et al. (2020); Karamanolakis et al. (2021); Yu et al. (2021). One typical issue of self-training with weak supervision is that the model tends to accumulate errors from noisy labels Guo et al. (2017), leading to deteriorated performance. To address this problem, Yu et al. (2021) adapts contrastive learning along with a set of reweighting and regularization strategies to suppress label noise in pseudo label generation, where weak labels are only used for model initialization. In contrast, our method makes full use of weak labels and performs weakly supervised learning via encouraging consistency between the predictor and the corrector, which helps to generate more reliable pseudo labels.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate our method on 8 textual datasets covering five text classification tasks from the WRENCH benchmark Zhang et al. (2021b), which is designed to make fair comparisons for weak supervision methods. The tested datasets include: IMDB Maas et al. (2011) and Yelp Zhang et al. (2015) for sentiment analysis, SMS Almeida et al. (2011) for spam classification, AGNews Zhang et al. (2015) for topic classification, TREC Voorhees & Tice (1999) for question classification, as well as Spouse Corney et al. (2016), CDR Davis et al. (2017), and SemEval Hendrickx et al. (2010) for relation extraction. The detailed statistics of these datasets are summarized in Appendix.

Training protocol. We perform weak supervision in a two-stage manner, where label models are first trained to generate weak labels and the end model is then trained with the generated weak labels for final predictions. We use RoBERTa-base Liu et al. (2019) as the end model and the pre-trained encoder for weak label modeling methods and self-training methods, respectively. In addition, we also test one-stage methods that learn the label model and the end model jointly.

Competing methods. We compare our SELCOR method against the following groups of competing methods. *True Label*: A fully-supervised baseline obtained by fine-tuning the RoBERTa-base model and a task-specific classification head with true labels, which serves as an upper bound of classification performance for weak supervision. *(Weighted) Majority Voting*: A simple and classical label modeling method for weak label generation. The weighted version further reweights the final votes by the label prior. *Data Programming* Ratner et al. (2016): It leans a generative probabilistic graphical model to recover the latent true labels from noisy weak supervision sources. *MeTaL* Ratner et al. (2019b): It models multi-source weak supervision with a Markov Network in the multi-task learning setting and estimates the true label by solving a matrix completion-style problem. One-stage and self-training based methods such as *Denoise* Ren et al. (2020): It adopts an attention network to estimate label reliability, and learns to aggregate weak labels along with the end model training. *COSINE* Yu et al. (2021): It generates and exploits pseudo labels to improve model generalization while applying contrastive learning and reweighting to suppress label noise caused during self-training.

Evaluation. For each dataset, we use the same performance metric in Zhang et al. (2021b) for fair comparisons. We repeat five runs of our method with different random seeds and report the average values and standard deviation of the performance metric on the test set. We adopt early stopping to select the best checkpoint on the validation set for final evaluation.

Implementation details. For our SELCOR model, the predictor h_{pred} , false positive corrector h_{fp} , false negative corrector h_{fn} are implemented by a linear projection into the C -dimensional label space followed by a softmax activation. We use AdamW (Loshchilov and Hutter, 2019) with a learning rate of $1e-5$ for optimization, and use grid search to determine the best hyper-parameters on the validation set. Specifically, we select the trade-off parameter γ from $\{0, 0.01, 0.1, 0.5, 1, 2\}$, the weak label smoothing parameter α from $\{0, 0.1, 0.2, 0.3, 0.4\}$, and the pseudo label update period T' from $\{10, 30, 50, 100, 200\}$. The detailed parameter configurations for each task are provided in Appendix. All experiments are conducted on a Nvidia GeForce RTX 3090 GPU.

4.2 RESULTS AND ANALYSIS

Table 1 shows the classification results for weakly-supervised learning methods, where the results of the competing methods are obtained from Zhang et al. (2021b). The results for True Label on Spouse are absent since there is no ground-truth label in Spouse for training. On all the datasets except AGNews, SELCOR achieves the best performance among weakly supervised learning methods and outperforms the second best method, COSINE, by 3.35% in the corresponding performance metric on average. This demonstrates that weak labels are in fact informative and should be fully utilized to improve the performance of weak supervision.

In AGNews, there exist many unlabeled data (around 28,800), which have been used to train COSINE. The superior performance of COSINE on AGNews over SELCOR could be attributed to the exploited information of unlabeled data. In comparison with other weakly supervised counterparts, self-training based methods such as COSINE and SELCOR show clear advantages, indicating the effectiveness of self-training in denoising weak labels. It is worth noting that our method is only

Table 1: Classification results on textual datasets from the WRENCH benchmark (**Best**).

Method	IMDb (Acc.)	Yelp (Acc.)	SMS (F1)	AGNews (Acc.)	TREC (Acc.)	Spouse (F1)	CDR (F1)	SemEval (Acc.)	Average
True Label	93.25 (0.30)	97.13 (0.26)	96.31 (0.58)	91.39 (0.38)	96.68 (0.82)	– –	65.86 (0.60)	93.23 (1.83)	90.55
Majority Voting	85.76 (0.70)	89.91 (1.76)	94.17 (2.88)	86.88 (0.98)	66.28 (1.21)	17.99 (1.99)	55.07 (3.47)	84.00 (0.84)	72.51
Weighted Majority Voting	86.06 (0.88)	82.27 (4.11)	92.96 (1.71)	86.70 (0.51)	58.88 (0.92)	16.14 (1.40)	42.37 (21.19)	67.47 (6.93)	66.61
Data Programming	86.26 (1.02)	89.59 (2.87)	28.25 (2.83)	86.81 (0.42)	72.12 (4.58)	17.62 (4.24)	54.42 (5.32)	70.57 (0.83)	63.21
MeTaL	84.98 (1.07)	89.08 (3.71)	93.28 (1.57)	87.18 (0.45)	60.04 (1.18)	16.42 (2.79)	53.68 (4.00)	70.73 (0.68)	69.42
Denoise	76.22 (0.37)	71.56 (15.80)	91.69 (1.42)	83.45 (0.11)	56.20 (6.73)	22.47 (7.50)	56.54 (0.37)	80.83 (1.31)	67.37
COSINE	88.22 (0.22)	94.23 (0.20)	96.67 (0.37)	88.15 (0.30)	77.96 (0.34)	40.50 (1.23)	60.38 (0.05)	86.20 (0.07)	79.04
SELCOR	89.31 (1.75)	95.78 (0.32)	97.44 (0.62)	87.35 (0.66)	79.92 (1.35)	54.46 (6.33)	68.23 (0.16)	86.63 (1.45)	82.39

trained on data with weak labels, while the competing method, COSINE, uses both weakly labeled and unlabeled data. Nevertheless, our method still outperforms COSINE by a large margin on average, indicating the effectiveness of the corrector in controlling and guiding the predictor to generate more reliable pseudo labels. As discussed in Sec. 5, we can readily obtain pseudo labels of unlabeled data from the predictor. We expect the performance of our method could be further improved by exploiting unlabeled data in a proper way.

4.3 EFFECT OF SELF-CORRECTION

In this section, we study how *self-correction* affects weak supervision. Figure 2 shows the training loss curves of SELCOR on the IMDb, Yelp, SMS, and TREC datasets. Pred and Corr present the training loss curves in terms of *pseudo labels* for the predictor and the corrector, respectively. Pred_gt and Corr_gt present the true loss curves in terms of *true labels* for the predictor and the corrector, respectively. Here, we provide Pred_gt and Corr_gt to indicate how well the model preserves the true label information during self-training, and our SELCOR model is still trained with pseudo labels. As can be seen, the curves for Pred_gt and Corr_gt are highly correlated. This means that the predictor and the corrector are highly correlated, and thus their losses should be made consistent in self-training.

Figure 2 shows the training loss curves of the predictor for SELCOR w/ and w/o self-correction. Self-train and SELCOR present the loss curves of the predictor for SELCOR w/ and w/o self-correction, respectively. Similarly, Selftrain_gt and SELCOR_gt present the true loss curves of the predictor for SELCOR w/ and w/o the corrector, respectively. As can be seen, although the training losses are still small, the true losses become large as the number of iterations increases. This implies that the model accumulates the errors in pseudo labels and suffers from overfitting. For all the datasets, the curves of SELCOR_gt are more stable and generally under those of Selftrain_gt. This demonstrates that by exploiting the consistency between label generation and label correction, self-correction is effective in alleviating error accumulation and thus leads to better performance.

4.4 EFFECT OF NOISE MODELS

The noise model for parameterizing the correcting posterior $p(y_{ji}|\tilde{y}_{ji}, \mathbf{x}_i)$ is the key component of our SELCOR model. In (2), we construct the noise model by combining the probabilities of a pseudo label being false positive and false negative. In this section, we study whether both the false positive and false negative components are necessary for modeling $p(y_{ji}|\tilde{y}_{ji}, \mathbf{x}_i)$. Table 2 provides the classification results for SELCOR with different noise models, where “FP only” and “FN only” represent the noise model constructed by the false positive component, i.e., $p(y_{ji}|\tilde{y}_{ji}, \mathbf{x}_i) =$

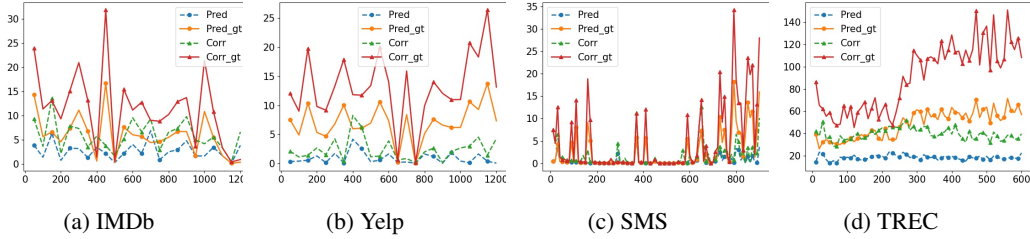


Figure 2: Loss curves of SELCOR over iterations.

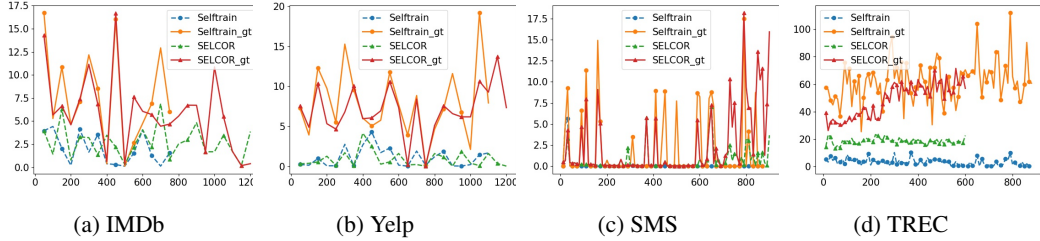


Figure 3: Loss curves for SELCOR w/ and w/o self-correction.

$(1 - p(y_{ji} = 0 | \tilde{y}_{ji} = 1, \mathbf{x}_i)) \cdot \tilde{y}_{ji}$, and the false negative component, i.e., $p(y_{ji} = 1 | \tilde{y}_{ji} = 0, \mathbf{x}_i) \cdot (1 - \tilde{y}_{ji})$, respectively. The experimental results show that failing to consider either false positive or false negative probabilities will result in degraded performance. This demonstrates that both the false positive and false negative components are important in estimating the label noise transition, and the proposed noise model (2) is effective in capturing the correlations between weak labels and pseudo labels for weakly supervised learning.

4.5 ABLATION STUDIES

The main components of our SELCOR model include the corrector, the predictor, parameter freezing for the predictor, and weak label smoothing. Table 3 provides the ablation study results for SELCOR on all 8 datasets. As a baseline, we also test the initialized SELCOR model without self-training, which is referred to as “Initialization”. The results of w/o predictor and w/o smoothing are the same as the full version of SELCOR. This is because the selected parameters γ and α are zeros, and thus the predictor and label smoothing have no effect on the performance. As can be seen, all the components contribute to the classification performance. In particular, when the corrector is removed from the SELCOR model, the performance drops the most by 2.62% on average. This suggests that the corrector contributes the most to the improved performance. The predictor is also important for SELCOR. Removing it leads to a drop of 2.39% on average. It is worth noting that without the corrector SELCOR fails to outperform the initialized model on IMDB and SMS. On the other hand, without the predictor SELCOR does not improve the initialized model on AGNews and CDR. These imply that the corrector and predictor capture complementary information and should be used together for the best performance. Despite its simplicity, freezing the predictor parameters consistently improves the performance. This can be attributed to the simplified model and more identifiable correlations between the predictor and the corrector. Finally, weak label smoothing is also useful to regularize the corrector by preventing it from underfitting.

5 DISCUSSION

Unlabeled data. One limitation of our method is that we do not consider unlabeled data that have no weak label for training. Such unlabeled data are available in many applications and can be used for better performance. Although the pseudo labels can be readily obtained by feeding unlabeled data

Table 2: Classification results of SELCOR with different noise models (**Best**).

Method	IMDb (Acc.)	Yelp (Acc.)	SMS (F1)	AGNews (Acc.)	TREC (Acc.)	Spouse (F1)	CDR (F1)	SemEval (Acc.)	Average
SELCOR	89.31 (1.75)	95.78 (0.32)	97.44 (0.62)	87.35 (0.66)	79.92 (1.35)	54.46 (6.33)	68.23 (0.16)	86.63 (1.45)	82.39
FP Only	87.08 (3.44)	94.99 (0.94)	97.25 (0.63)	87.14 (1.16)	71.12 (5.45)	53.28 (5.69)	68.52 (0.69)	85.57 (1.31)	80.62
FN Only	86.52 (3.07)	94.91 (0.91)	97.43 (0.63)	86.94 (0.89)	75.12 (7.76)	51.24 (5.24)	67.90 (0.48)	85.93 (1.53)	80.75

Table 3: Ablation study results for SELCOR (**Best**).

Method	IMDb (Acc.)	Yelp (Acc.)	SMS (F1)	AGNews (Acc.)	TREC (Acc.)	Spouse (F1)	CDR (F1)	SemEval (Acc.)	Average
Initialization	84.94 (1.26)	88.94 (1.98)	96.90 (0.25)	87.07 (0.48)	67.56 (0.55)	48.18 (0.77)	67.70 (0.31)	84.43 (0.30)	77.32
SELCOR	89.31 (1.75)	95.78 (0.32)	97.44 (0.62)	87.35 (0.66)	79.92 (1.35)	54.46 (6.33)	68.23 (0.16)	86.63 (1.45)	82.39
w/o Corrector	84.94 (1.26)	95.04 (0.32)	96.90 (0.25)	87.19 (0.50)	72.80 (3.72)	53.09 (2.53)	68.40 (0.28)	86.17 (0.67)	79.77
w/o Predictor	86.94 (1.59)	93.87 (0.71)	97.17 (0.20)	86.91 (0.46)	76.40 (3.54)	54.46 (6.3)	67.85 (0.14)	84.60 (0.31)	80.51
w/o Freezing	88.61 (0.53)	94.76 (0.38)	97.34 (0.30)	87.07 (0.48)	76.24 (3.61)	50.41 (1.36)	68.04 (0.07)	85.83 (0.53)	80.35
w/o Smoothing	88.43 (1.59)	95.61 (0.20)	97.10 (0.40)	87.06 (0.52)	75.04 (2.59)	54.46 (6.3)	68.30 (0.14)	87.00 (0.52)	80.85

into the predictor, it is unclear how to incorporate such label information in the corrector without knowing the corresponding weak labels. We will leave it as future work.

Further extensions. In our experiments, we focus on NLP tasks while our method is general and could be applied to tasks beyond NLP if the encoder has enough capacity for self-training. For instance, it would be interesting to test our method for image classification with crowd-sourcing labels. In addition, our method has the potential to be further boosted by adapting existing techniques such as confidence-based reweighting, entropy regularization, and contrastive learning, which have shown to be effective in weakly supervised learning. As a key component of the proposed method, we construct a noise model to learn the noise transition in the binary classification setting. It is possible to develop a more general and sophisticated noise model that connects weak labels to pseudo labels with more flexibility.

6 CONCLUSION

In this paper, we propose SELCOR, a self-training based method for weakly supervised learning without using any ground-truth label. SELCOR aims to jointly learn a predictor and a corrector upon a shared encoder, where the predictor generates pseudo clean labels and the corrector revises weak labels to *reproduces* the pseudo labels generated by the predictor. By encouraging consistency between the predictor and the corrector, SELCOR prevents the predictor from generating suspicious pseudo labels that significantly mismatch the noise model defined in the corrector, and enables the predictor and the corrector iteratively improve each other in self-training. Experimental results on 8 benchmark datasets show that our method takes advantage of exploiting the consistency between label generation and label correction in reducing label noise, and outperforms existing weakly supervised methods by large margins.

REFERENCES

- Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *DocEng*, pp. 259–262, 2011.
- Chidubem Arachie and Bert Huang. A general framework for adversarial label learning. *JMLR*, 22: 1–33, 2021.
- Chidubem Arachie and Bert Huang. Data consistency for weakly supervised learning. *arXiv preprint arXiv:2202.03987*, 2022.
- Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, pp. 233–242, 2017.
- Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. Learning from rules generalizing labeled exemplars. In *ICLR*, 2020.
- Baixu Chen, Jinguang Jiang, Ximei Wang, Jianmin Wang, and Mingsheng Long. Debaised pseudo labeling in self-training. *arXiv preprint arXiv:2202.07136*, 2022.
- Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *Proceedings of ICML*, pp. 1789–1799, 2020.
- D. Corney, M. Albakour, Miguel Martinez-Alvarez, and Samir Moussa. What do a million news articles look like? In *NewsIR@ECIR*, 2016.
- Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of WWW*, pp. 285–294, 2013.
- Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Benjamin L King, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45:D972–D978, 2017.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of Machine Learning Research*, volume 70, pp. 1321–1330. PMLR, 2017. URL <http://proceedings.mlr.press/v70/guo17a.html>.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Semeval*, pp. 33–38, 2010.
- Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in NeurIPS*, 31, 2018.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL*, pp. 541–550, 2011.
- Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Weakly supervised image classification through noise regularization. In *Proceedings of CVPR*, June 2019.
- Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *Proceedings of ACL*, pp. 2410–2420, 2016. doi: 10.18653/v1/P16-1228. URL <https://www.aclweb.org/anthology/P16-1228>.
- Giannis Karamanolakis, Subhabrata (Subho) Mukherjee, Guoqing Zheng, and Ahmed H. Awadallah. Self-training with weak supervision. In *NAACL*, 2021.
- David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Proceedings of NeurIPS*, volume 24, pp. 1953–1961, 2011.
- Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *ICLR*, 2018.

- Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 2, 2013. URL https://www.kaggle.com/blobs/download/forum-message-attachment-files/746/pseudo_label_final.pdf.
- Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of SIGKDD*, pp. 1054–1064, 2020. ISBN 9781450379984. URL <https://doi.org/10.1145/3394486.3403149>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.
- Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of ACL*, pp. 430–439, 2017. doi: 10.18653/v1/P17-1040. URL <https://www.aclweb.org/anthology/P17-1040>.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of ACL*, pp. 142–150, 2011.
- Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H Bach, and Eli Upfal. Adversarial multi class learning under weak supervision with performance guarantees. In *Proceedings of ICML*, pp. 7534–7543, 2021.
- Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of ICDM*, pp. 86–93, 2000.
- Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of CVPR*, 2017.
- Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 05 2016.
- Alexander Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29, 07 2019a. doi: 10.1007/s00778-019-00552-1.
- Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. *Proceedings of AAAI*, 33: 4763–4771, 07 2019b. doi: 10.1609/aaai.v33i01.33014763.
- Mengye Ren, Wenyan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4334–4343. PMLR, 10–15 Jul 2018. URL <http://proceedings.mlr.press/v80/ren18a.html>.
- Wendi Ren, Yinghao Li, Hanling Su, David Kartchner, Cassie Mitchell, and Chao Zhang. Denoising multi-source weak supervision for neural text classification. In *Findings of EMNLP*, pp. 3739–3754, 2020. URL <https://www.aclweb.org/anthology/2020.findings-emnlp.334>.
- Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-end weak supervision. In *Proceedings of NeurIPS*, volume 34, pp. 1845–1857, 2021.
- Esteban Safranchik, Shiyang Luo, and Stephen H Bach. Weakly supervised sequence tagging from noisy rules. In *Proceedings of AAAI*, 2020.
- Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in NeurIPS*, 32, 2019.

- Ellen M. Voorhees and Dawn M. Tice. The TREC-8 question answering track evaluation. In Ellen M. Voorhees and Donna K. Harman (eds.), *Proceedings of The Eighth Text REtrieval Conference*, volume 500-246, 1999. URL <http://trec.nist.gov/pubs/trec8/papers/qa8.pdf>.
- Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. Learning with noisy labels for sentence-level sentiment classification. In *Proceedings of EMNLP*, pp. 6286–6292, 2019. doi: 10.18653/v1/D19-1655. URL <https://www.aclweb.org/anthology/D19-1655>.
- Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of ICML*, pp. 478–487, 20–22 Jun 2016. URL <http://proceedings.mlr.press/v48/xieb16.html>.
- Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv preprint arXiv:2203.09081*, 2022.
- David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL*, pp. 189–196, Cambridge, Massachusetts, USA, 1995. doi: 10.3115/981658.981684. URL <https://aclanthology.org/P95-1026>.
- Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of NAACL*, pp. 1063–1077, 2021.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.
- Jieyu Zhang, Yue Yu, , Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. Wrench: A comprehensive benchmark for weak supervision. In J. Vanschoren and S. Yeung (eds.), *Proceedings of NeurIPS*, volume 1, 2021b.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of NeurIPS*, pp. 649–657, 2015. URL <http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf>.
- Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8778–8788. Curran Associates, Inc., 2018.
- Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proceedings of AAAI*, volume 35, pp. 11053–11061, 2021.

Table 4: Statistics of the tested datasets.

Dataset	Task	# Class	# Train	# Dev	# Test	% Coverage	% Accuracy
IMDB	Sentiment	2	20,000	2,500	2,500	87.58	69.88
Yelp	Sentiment	2	30,400	3,800	3,800	82.78	73.05
SMS	Spam	2	4,571	500	2,719	40.52	97.26
AGNews	Topic	4	96,000	12,000	12,000	69.08	81.66
TREC	Question	6	4,965	500	500	95.13	75.92
Spouse	Relation	2	22,254	2,811	2,701	25.77	-
CDR	Relation	2	8,430	920	4,673	90.72	75.27
SemEval	Relation	9	1,749	200	692	100.00	97.69

Table 5: Parameter configurations of SELCOR on different datasets.

Dataset	IMDB	Yelp	SMS	AGNews	TREC	Spouse	CDR	SemEval
Batch size	16	16	16	16	32	16	16	16
Max #tokens	256	512	256	256	256	512	512	512
T'	50	30	50	100	10	100	200	30
α	0.4	0.2	0.4	0.4	0.1	0.2	0.2	0.1
γ	0.5	2	0.5	2	0.01	0.01	1	0.1

A APPENDIX

A.1 DATASET STATISTICS

Table 4 summarizes the statistics of the datasets tested in our experiments.

A.2 PARAMETER CONFIGURATIONS

Table 5 provides the parameter configurations of our SELCOR model on different datasets. The code of our SELCOR method is available at <https://anonymous.4open.science/r/SELCOR>.