Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrixes

Anonymous ACL submission

Abstract

Distant supervision significantly reduces human efforts in constructing training data for many classification tasks, however, it inevitably introduces noise into the dataset. In this paper, we take a deep look at the application of distant supervision in relation extraction. We investigate the underlying noise in the training data, and propose to utilize constrained dynamic transition matrices to model the noise. We also take the advantage of curriculum learning to explore possible clues of the noise patterns, and further improve the model robustness. We experiment our models with both sentence level and bag level in the situations with and without prior knowledge of data quality. The experimental results show that our method can better handle the noise in distantly supervised relation extraction and improve extraction performance in various settings.

1 Introduction

TODO: ZW: Start by saying DS is a powerful technique for performing relation extraction on a large text corpus where manually labeling every single sentence is impossible. (2) The quality of DS highly depends on the training examples, but it not always possible to have a perfectly labelled data set and noise is pervasive in training data. Therefore, DS-based RE has to work with noisy data. (3)There have been attempts on modelling/reducing the noises in the training data set, but prior work are terrible... (4) We then move to talk about our key insights, how fantastic our approach is and how does it go beyond the state-of-the-art. (5) Summarise our experimental results, saying that we beat the state-of-the-art by xx%.

Distant supervision is a way of exploiting prior knowledge to construct training data for large scale classification tasks, without relying on laborious human annotation. The basic paradigm lies in making proper assumptions according to the nature of a task, distilling simple yet effective rules from prior knowledge and finally applying these rules to automatically prepare training data. Successful distant supervision applications include, relation extraction (Mintz et al., 2009), cross-lingual semantic analysis (Fang and Cohn, 2016), and so on. The former assumes that any sentence containing both the subject and object of a $< subj, rel, obj >$ knowledge triple can be seen as a support for the existence of relation rel between $subj$ and obj. Therefore, one can easily create large scale training data by aligning the triples in a structured knowledge base with free text automatically, without manually annotating a sentence.

Although distant supervision can significantly reduce the cost of obtaining training data, it also inevitably brings noise to the data, since the assumptions may not be perfect. TODO: F: we need example here. Shall we put the at-least-one assumption here? or somewhere later? We find that many automatically labeled-positive data are actually negative instances (*false positive*), while some labeled-negative data are real positive instances (*false negative*). Furthermore, there may also be confusions between positive labels. All these noisy data inevitably affect the training procedure, and lead to a classification model with inaccurate performances. TODO: F: there have been several work addressing the noises, but ..... for example, *at-least-one assumption*. we may talk about sentence level and bag level here.

We need to note that, the noise introduced by distant supervision is not random, and the resulting dataset may speak for itself, e.g., containing useful clues to identify its noise pattern. Take dis-

tantly supervised relation extraction as an example, many sentences automatically labeled to express the relation *place_of_birth* between a person and a place may only talk about the work place of this person. F: this is the motivation for TM, right? but it is hard to understand, we need something simple but easy to get the point. Since we also find many sentences labeled as *place_lived* talks about the person's work place, we can reasonably assume that if a sentence is talking about the work place of a person, although the real relation expressed by the sentence is *place_lived*, there are still chances that it is erroneously labeled as *place_of_birth* by distant supervision.

F: We need to say, TM is a way to model the noise, better than previous work? but still need ways to control the noise though trace regularization???? This shows that it is possible to identify the noise pattern by analyzing the input data. In this paper, we propose to dynamically generate a transition matrix $\mathbf{T}$ for each datum to model the transition from the true label to the observed label. Here T_{ij} represents the conditional probability that the observed label is j given the true label is i. Since the label modeled by transition matrix can be both positive and negative, this method actually has the ability to model all the three types of noise introduced by distant supervision. F: i do not know why we need this sentence? The dynamically built transition matrices provide us the oppurtunity to characterize the noise of the data, which should be properly treated to control the noise.

F: Here we need to gently shift our focus from noises to quality of the data, and combine TM and curriculum learning. In some cases, the rules used to build the training data may be further exploited to identify the noise level inside the data. For example, ...

F: we do not need this, but we need how TM works to model the noise and how regularization helps to control the effect of noise controlling. In this paper, we focus specifically on the relation extraction task. The data is noisy because not all sentences containing $subj$ and obj support the ($subj$, rel, obj) triple. TODO: for example... In previous literature, this noise is often implicitly handled by the *at-least-one assumption* that at least one of the sentences containing both $subj$ and obj support the ($subj$, rel, obj) triple. The sentences that containing both $subj$ and obj are therefore aggregated into a sentence bag, and the problem becomes classifying the relation expressed by the sentence bag instead (Riedel et al., 2010; Lin et al., 2016).

F: we do not need this, but how to find the clues about data quality and combine them into a CL framework, and work with Regularized TM to improve. However, the at-least-one assumption is not perfect either, and therefore introduces bag level noise. First, if all the retrieved sentences do not support the ($subj$, rel, obj) triple, this bag is false positive. Second, if the triple is true but missing in the KB, then the bag is false negative.

We apply our transition matrix method to both sentence level and bag level models. We also propose to combine curriculum learning and trace regularization for training to deal with the lack of direct guidance over the noise. The experiments show that our transition matrix method improves both of these models in two datasets. We also show that our training procedure can make use of the prior knowledge of data quality and help the transition matrix model the noise better.

2 Motivation

TODO: ZW: We need a concrete example here to demonstrate what kind of noise can exist in the data and their impact on relation extraction. We then use this example to explain our approach. F: this could be put into the introduction, saying different types of noises in the data, and all thoses noises can be modeled by TM.

As a motivation example, consider xx...

This example shows that noisy data can have a significant impact on relation extraction. What we like to have is a technique that can xxx. In the remainder of this paper, we describe such an approach based on xxx.

3 Task Description

we need formal description for the task. Distantly supervised relation extraction aims at extracting $< subj, rel, obj >$ triples from free text using distantly obtained training data, which can be examined in two different settings, i.e., sentence level, and bag level. The former takes a sentence s containing both $subj$ and obj as input, and output the relation expressed by the sentence between $subj$ and obj. The latter setting is based on the *at-least-one* assumption and takes a bag of sentences S as input where each sentence $s \in S$ contains both

$subj$ and obj, and output the relation expressed by the sentence bag between $subj$ and obj.

In this paper, we will inject our noise modeling approach in both settings, and further investigate the performance of our approach in the situation with and without prior knowledge of the data quality. Specifically, we assume the prior knowledge can help us roughly distinguish reliable data from unreliable ones, and therefore split the dataset into several subsets with different levels of reliability. If no prior knowledge can be used, all the data are treated equally.

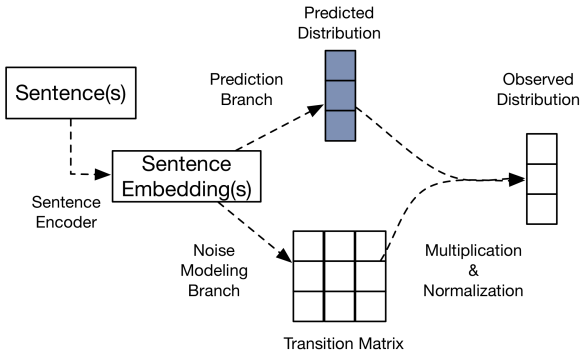


Figure 1: The architecture of our denoising model

4 Our Approach

The framework of our model is shown in Figure 1. First, the input sentence (or sentence bag) is passed to a sentence encoder to get sentence embedding(s). After that, the model is split into two branches. The prediction branch generates the predicted relation distribution $\mathbf{p}$ of the input sentence (or sentence bag). The noise modeling branch generates the transition matrix $\mathbf{T}$. Finally, the predicted distribution is multiplied by the transition distribution to generate the observed relation distribution $\mathbf{o}$. The predicted relation distribution $\mathbf{p}$ is the output of the model while the observed relation distribution $\mathbf{o}$ is used to simulate the relation assigned by distant supervision. In this way, the noise is modeled by the transition matrix and the real prediction is protected from the influence of the noise. In rest of this section, we will first describe the sentence level model, and then extend the model to bag level.

4.1 Sentence Encoder

The sentence encoder serves to transform an input sentence to an embedding vector that encodes semantic meaning of the sentence. Theoretically, almost any sentence encoder would work here. Similar to previous researches, we also use the piecewise convolutional neural network (PCNN) model (Zeng et al., 2015) as our sentence encoder. First, the distances of each word to the subject and the object are embedded as randomly initialized vectors and concatenated to the original word vector. After that, the PCNN model divides the input sentence into three pieces by the subject and the object, and apply convolutional neural network (CNN) to each piece to calculate the piece embedding. The final sentence embedding is the concatenation of the embeddings of the three pieces.

4.2 Prediction Branch

The prediction branch generates the predicted relation distribution $\mathbf{p}$ and can be implemented by the prediction part of almost any relation extraction neural network models. Similar to (Luo et al., 2016), for sentence level models, we first feed the sentence embedding to a full connection layer, and then use softmax for relation classification.

4.3 Noise Modeling Branch

The noise modeling branch calculates a transition matrix dynamically for each sentence (or sentence bag) to model its noise pattern.

For sentence level models, the sentence embedding $\mathbf{x}$ is passed to another full connection layer to obtain the sentence embedding $\mathbf{x}_n$ used specifically for noise modeling branch. The transition matrix $\mathbf{T}$ is calculated using softmax function :

$$T_{ij} = \frac{exp(\mathbf{w}_{ij}^T \mathbf{x}_n + b)}{\sum_{j=1}^{|\mathbb{C}|} exp(\mathbf{w}_{ij}^T \mathbf{x}_n + b)} \quad (1)$$

where T_{ij} is the conditional probability that this sentence is labeled as relation j by distant supervision given i as the true relation, $\mathbf{w}_{ij}$ is the weight vector for this situation, b is a scalar bias and $|\mathbb{C}|$ is the number of relations. Note that the softmax function guarantees that each row of the transition matrix $\mathbf{T}$ sums to 1.

4.4 Observed Relation Distribution

The observed relation distribution $\mathbf{o}$ is calculated by multiplying transition matrix T and prediction probability P: .

$$\mathbf{o} = \mathbf{T}^T \cdot \mathbf{p} \quad (2)$$

where $\cdot$ represents dot product and we normalizes the elements of $\mathbf{o}$ so that $\sum_i o_i = 1$ afterwards.

Different from previous works that use the predicted relation distribution $\mathbf{p}$ to directly match the relation labeled by distant supervision. We instead want??? $\mathbf{p}$ to match??? the true relation and use $\mathbf{o}$ to match the noisy label. Note that if the true relation of the input training instance is i, we can assume that this relation i could be labeled as relation j by distant supervision with probability T_{ij}. Therefore, Equation 2 actually models the procedure of how the noisy label is produced and thus can help the model make better use of the noisy training data. How to better use?? by using regularization????

Also note that the noise modeling branch and $\mathbf{o}$ is only used in the training phase. In the test phase, we only use the prediction brach and take the predicted relation distribution $\mathbf{p}$ as our output. TODO: last paragraph can be removed

4.5 Bag Level Models

Bag Level Prediction Branch The key problem in the bag level prediction branch is how to aggregate the embeddings of each sentence in the bag. Here we experiment with two methods, average and attention methods. The average aggregation calculates the bag embedding $\mathbf{s}$ by averaging the embeddings of each sentence, and the resultant bag embedding is fed to a softmax classifier for relation classification.

The attention aggregation method is proposed by (Lin et al., 2016). It calculates an attention value for each sentence with respect to each relation, and uses the following equation to calculate the bag embedding with respect to relation j:

$$\mathbf{s}_j = \sum_i^n \alpha_{ij} \mathbf{x}_i \tag{3}$$

where $\mathbf{x}_i$ is the embedding of sentence i, n is the number of sentences in the bag and α_{ij} is the attention value over sentence i with respect to relation j. The resultant bag embedding is fed to a softmax classifier to predict the probability of relation j.

Bag Level Transition Matrix THis is only for attention aggregation method?? how about average aggregation one?? Here we also use the attention mechanism to calculate the bag embedding with respect to each relation with Equation 3, and the attention value for sentence i with respect to relation j is calculated by:

$$\alpha_{ij} = \frac{exp(\mathbf{x}_i^T \mathbf{r}_t^j)}{\sum_i^n exp(\mathbf{x}_i^T \mathbf{r}_t^j)} \tag{4}$$

where $\mathbf{x}_i$ is the embedding of sentence i and $\mathbf{r}_t^j$ is the randomly initialized embedding of relation j used specifically for noise modeling branch.

Then the transition matrix $\mathbf{T}$ is calculated by:

$$T_{ij} = \frac{exp(\mathbf{s}_i^T \mathbf{r}_t^j + b)}{\sum_{j=1}^{|\mathbb{C}|} exp(\mathbf{s}_i^T \mathbf{r}_t^j + b)} \tag{5}$$

where $\mathbf{s}_j$ is the bag embedding with respect to relation j, $\mathbf{r}_t^j$ is the same embedding of relation j.

5 Training Procedure

I would suggest to move some of the motivations to the Introduction part, we need to motivate regularization and CL there.The difficulty of training the transition matrix lies in that there is no direct guidance over the noise pattern. The only supervision we have is the noisy label produced by distant supervision. However, if we have prior knowledge to roughly separate the data into reliable and unreliable parts, we can use the split as indirect supervision over the noise pattern by letting the model treat these data differently. In this section, we describe how to use trace regularization to control the behavior of the transition matrix. Furthermore, instead of modeling the noise at the very beginning of the training, we emphasis the noise modeling gradually by building different curriculums in the situation with and without prior knowledge of the data quality under the curriculum learning framework. Apart from that, we also show how to constrain the ability?? of the transition matrix to avoid overfitting???.

5.1 Trace Regularization

Intuitively, if the noise is small, the transition matrix $\mathbf{T}$ will tend to become an identity matrix (vice versa). Since each row of $\mathbf{T}$ sums to 1, the similarity between the transition matrix and the identity matrix can be represented by the trace of the transition matrix $\mathbf{T}$. The larger the $trace(\mathbf{T})$ is, the smaller the elements that do not lie in the diagonal are, and the more similar the transition matrix $\mathbf{T}$ is to identity matrix. Therefore, we can realize our expectation over the noise level of the data by controlling the value of $trace(\mathbf{T})$.

5.2 Curriculum Learning

The basic idea of curriculum learning is simple: start with the easiest aspect of a task, and level up the difficulty gradually.

With Prior Knowledge of Data Quality If we have prior knowledge about which part of the training data is more reliable and which is unreliable, the most straightforward way to build a curriculum is by controlling reliability of ???? the training data. Specifically, we can first train the prediction branch on the reliable data for some epochs and then add the unreliable data as well as the noise modeling branch???. In this way, the prediction branch is roughly trained before exposed to more noisy data.

Furthermore, we can also utilize our prior knowledge of the data quality in the form of trace regularization. Specifically, we use cross entropy as basic loss function and the final loss function with trace regularization is: this sentence is weird!

$$Loss = \sum_{i=1}^{M}\sum_{j=1}^{N_i} -log(o_{ijy_{ij}}) + \beta_i trace(\mathbf{T}_{ij}) \tag{6}$$

where i is the index of the data subsets, j is the index of training data, β_i is the trace regularization weight for subset i, $\mathbf{T}_{ij}$, y_{ij} and $o_{ijy_{ij}}$ are the transition matrix, relation labeled by distant supervision, and the observed probability of that relation for datum j in subset i respectively.

For the reliable subset, we want $trace(\mathbf{T})$ to be large (negative β) so that the element values of $\mathbf{T}$ will be centralized to the diagonal and the transition matrix will be similar to identity matrix. As for the unreliable subsets, we want the $trace(\mathbf{T})$ to be small (positive β) so that the element values of their transition matrices will be diffusive and the transition matrix is encouraged to model the noise??? how and why???. Note that this loss function only works for sentence level models, since reliable sentences and unreliable ones are all aggregated into a sentence bag in the bag level models and therefore we can not determine which bag is reliable and which is not. However, bag level models can still use the curriculum by changing the content of the bag. so ??

Without Prior Knowledge of Data Quality If we do not have the prior knowledge about the quality of training data, we can still build a curriculum, which can be used in all situations, by controlling the training objectivegradually controlling the impact of the transition matrix. Specifically, we design a decreasing weighting scheme for the trace regularization component, defined as:

$$Loss = \sum_{i=1}^{N} -(\alpha log(o_{iy_i}) + (1-\alpha) log(p_{iy_i})) + \beta trace(\mathbf{T}_i) \tag{7}$$

where $0 \leq \alpha \leq 1$, y_i is the relation assigned by distant supervision for datum i, o_{iy_i} and p_{iy_i} are the probabilities that the observed and predicted relation for datum i is y_i respectively. Instead of only using the observed relation distribution $\mathbf{o}$ to simulate the relation labeled by distant supervision, we use the linear combination of the cross entropy of both the observed relation distribution $\mathbf{o}$ and the predicted relation distribution $\mathbf{p}$.

At the beginning of training, we set $\alpha = 1$ and $\beta < 0$, which means we do not expect to model the noise (easy part of the problem????). As the training proceeds, the prediction branch gradually learns the basic prediction ability, then we increase the difficulty level by decreasing α and the absolute value of β by ρ every τ epochs to gradually guide our model to learn to model the noise.

5.3 Constrained Transition Matrix

Recall that the bag level noise mainly consists of false negative and false positive why there is no label confusion in bag level??, but our transition matrix also has the ability to model the confusion among positive relations. To prevent overfitting and make the model concentrate on the false negative and false positive noise??, we restrict the transition matrix for bag level models so that only the diagonal, the first column and the first row of the transition matrix do not equal to zero (assume the index of *no-relation* is 0).

6 Evaluation

To explore the behavior of our transition matrix method under different settings, we experiment our transition matrix method with both sentence level and bag level models in the situation with and without prior knowledge of the data quality. We show that our method works in all of these settings and prior knowledge of the data quality can benefit the training of transition matrix. We also

find that the sentence level models works better when we have both reliable and unreliable data, but the bag level model performs better if all the data are treated equally. Furthermore, to explore the generalization ability of our method, we also conduct experiments in two datasets.

6.1 Data Set

The first dataset is proposed by (Luo et al., 2016) and aims at extraction relations between entity and time (*time RE data*). The dataset is constructed by aligning Wikidata triples with Wikipedia corpus. Based on the granularity of the time expression in the sentence, this dataset can be split into 3 subsets with different levels of reliability. The reliable subset is used as basic training data, validation data and test data which contains 22,214, 2,776 and 2,771 positive sentences respectively. The two less reliable subset contains 2,094 and 53,469 positive sentences and are used as additional training data. Negative data are constructed with two heuristic strategies. We use this dataset because this is a public dataset on relation extraction that has both reliable and unreliable data, which is suitable for all of our experiment settings.

We also experiment on the dataset proposed by (Riedel et al., 2010), which is a commonly used dataset in relation extraction (*entity RE*). This dataset is generated by aligning triples in Freebase with the New York Times corpus (NYT corpus). The training data contains 522,611 sentences and 281,270 entity pairs. The test set contains 172,448 sentences and 96,678 entity pairs. We experiment our bag level models in this dataset to see the generalization ability of our transition matrix model.

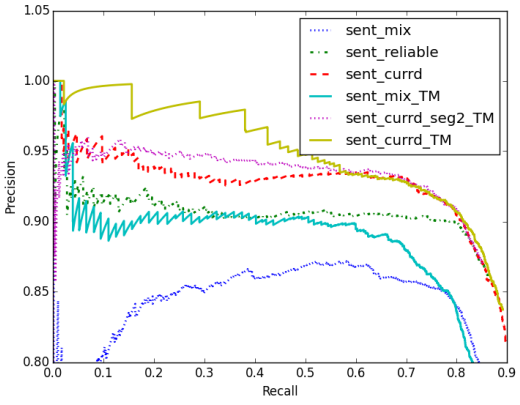


Figure 2: Sentence Level Results on Time RE

6.2 Hyper-parameters

Sentence Level Model We experiment our sentence level model on time RE data. We use 100-dimensional word embedding pre-trained using GloVe (Pennington et al., 2014) on Wikipedia and Gigaword, and 20-dimensional vector for distance embedding. The convolution window is 3 and the number of convolution kernels is 200. The size of the full connection layer is also 200. As for training, we use stochastic gradient descend (SGD) with batch size 20, learning rate 0.1. We also use dropout with probability 0.5 upon the sentence embedding. Each data subset is added after 15 epochs since the precious one is added. The trace regularization parameters for three subsets are $\beta_1 = -0.01$, $\beta_2 = 0.01$ and $\beta_3 = 0.1$ respectively from the reliable one to the most unreliable one (the ratio of β_3 and β_2 is fixed to 10 and 5 when tuning hyper-parameters).

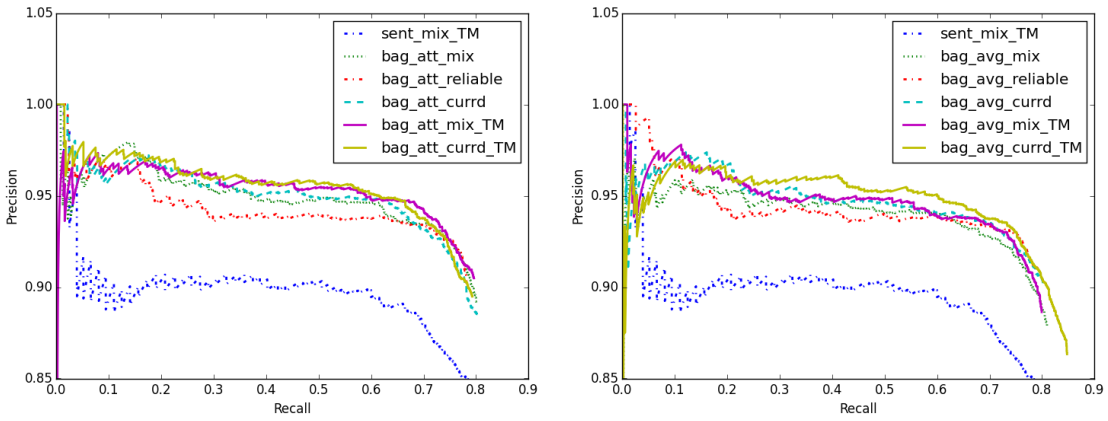
Bag Level Model The parameters of the bag level model is almost the same as the sentence level model on time RE data, except that the learning rate is 0.01. As for entity RE data, The word embedding is of dimension 50 and is pre-trained on the NYT corpus using word2vec[1]. The convolution window is 3 and the number of convolution kernels is 256, distance embedding size is 5, batch size is 16 and learning rate is 0.01. For all the bag level models, the linear combination parameter α is 1 and trace regularization parameter β is -0.1 at the start of training. We experiment with decay rate $\{0.95, 0.9, 0.8\}$ and decay step $\{3, 5, 8\}$. We find that using decay rate 0.9 and decay step 5 performs well in most situations.

6.3 Results on Time RE Data

Sentence Level Models The results of sentence level models are shown in Figure 2 in the form of precision recall curves (PR curves). We can see that the performance of the model trained on all subsets mixed together (*sent_mix*) is very bad, and is significantly worse than the model trained only on the reliable subset (*sent_reliable*), which shows that the noise problem is nonnegligible. However, with the help of transition matrix, the model obtains the ability of modeling noise (*sent_mix_TM*), which significantly improves the performance of the model. By using the reliable subset first and gradually adding less reliable data (*sent_currd*[2]),

[1] https://code.google.com/p/word2vec/

[2] *curr* refers to curriculum learning and *d* refers to data



(a) Attention Aggregation (b) Average Aggregation

Figure 3: Bag Level Results on Time RE

we can see that the model can actually make use of the noisy data and performs better than the model trained only on the reliable subset. However, by mixing the two unreliable subsets together and use the curriculum of training on the reliable data first as well as the transition matrix, we can see that the prior knowledge of data quality can help the transition matrix model the noise better and further improve the model performance. Furthermore, by using the curriculum of 3 subsets (*sent_currd_TM*) instead of 2, the performance can be further improved. Therefore, we will use 3 subsets in the rest of the experiments.

Bag Level Attention Aggregation Models The results of the bag level models with attention aggregation is shown in Figure 3(a). We can see that the basic bag level attention aggregation model (*bag_att_mix*) performs good and significantly outperforms the *sent_mix_TM* model. Recall that the bag level model is based on the *at-least-one assumption* that at least one of the sentences in the sentence bag support the ($subj$, rel, obj) triple, and the *sent_mix_TM* model do not use any assumption about the dataset. This shows that prior knowledge of the data quality plays an important role in the situation where the dataset is noisy. We can see that the model trained only on the reliable subset (*bag_att_reliable*) performs worse than the model trained on the mixed dataset, which shows that the attention aggregation has some denoising ability and can make use of the noisy data. However, since attention aggregation already has reasonable denoising ability, we can see that although adding noisy data gradually alone works well in sentence level model, it does not improve the attention aggregation model (*bag_att_currd*). Note that the at-least-one assumption does not always hold and there are also false negative and false positive problems in bag level. Therefore, we can see that using transition matrix with or without curriculum learning over the dataset (*bag_att_currd_TM* and *bag_att_mix_TM*) all improve the model performance, and *bag_att_currd_TM* performs slightly better.

Bag Level Average Aggregation Models The results of the bag level models with average aggregation is shown in Figure 3(b). The ranking of each setting is similar to the attention aggregation models. However, since its denoising ability is not as good as attention aggregation, adding unreliable data gradually (*bag_avg_currd*) improves the model performance here. We can also see that the transition matrix improves the average aggregation models more significantly than the attention aggregation models. Note that due to the inferior denoising ability of average aggregation, the unhandled sentence level noise may further propagates to bag level, which gives the transition matrix more chance to help model the noise.

Comparison The comparison of the best settings of each model family is shown in Figure 4. We can see that all of our transition matrix models outperform the model of (Luo et al., 2016). With the help of transition matrix, although the basic version of average aggregation is not as good as attention aggregation, its transition matrix version is similar to the attention aggregation. Also note

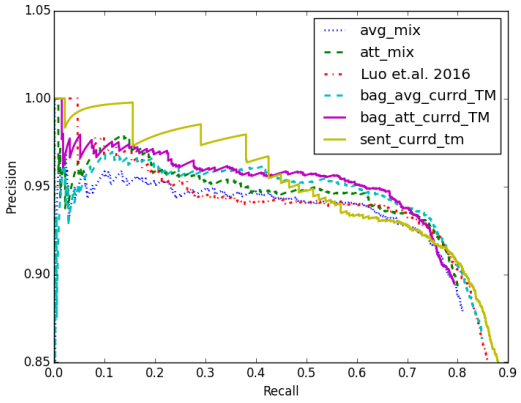


Figure 4: Comparison on Time RE

that although the sentence level models trained on mixed data do not perform very good, the sentence level model can use transition matrix to model the sentence level noise and thus performs best in all these models. Recall that the transition matrix can model the noise rather than just reduce the influence of noisy sentences as in bag level models, the sentence level model actually has the ability to make use of the noisy data. This shows that sentence level noise is more significant than the bag level noise in relation extraction, and modeling noise works better than just trying to reducing the influence of noise.

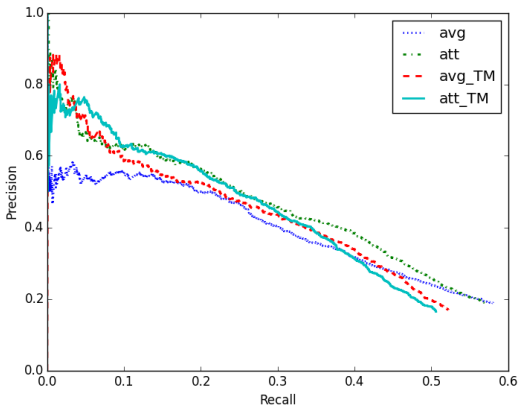


Figure 5: Results on Dataset of Riedel et.al.

6.4 Performance on Entity RE Data

To show the generalization ability of our proposed transition matrix method, we also conduct experiments on the entity RE dataset proposed by (Riedel et al., 2010). We implement the average aggregation method (*avg*) and the attention aggregation method (*att*) proposed by (Lin et al., 2016) as well as their corresponding transition matrix versions (*avg_TM* and *att_TM*). The results are shown in Figure 5. We can see that, due to the inferior denoising ability of average aggregation, it performs significantly worse than attention aggregation. Similar to the results in time RE data, since the unhandled sentence level noise propagates to the bag level, which makes the bag level noise become more severe, the transition matrix has more chance to model the noise. Therefore, the *avg_TM* model clearly outperforms the *avg* model. As for attention aggregation, this model already have good ability in reducing the impact of sentence level noise. Since the bag level noise is less significant than the sentence level noise, the improvement of our transition matrix model is limited, which only improves the model on the low recall part. Note that the low recall part corresponds to high precision, which is more useful than the rest of the extraction results in practice. Therefore, our transition matrix method is also useful in this situation.

7 Related Work

TODO: add more description about DS TODO: ZW: Starting by saying something like: Our work lies at the intersection of numerous areas: relation extraction, xx, xx. There is no work similar to ours, with respect to modeling the noise of training data for distant supervision and then use the knowledge of noise to improve relation extraction. We then organise the related work into different areas.

Distant Supervision Due to the simplicity and the good generalization ability of distant supervision, this method has been utilized in many tasks. (Go et al., 2009) builds training data for sentiment classification by using emoticons to label the sentiment polarity of sentences. (Ritter et al., 2011) constructs entity type classification data by aligning entity mentions with their possible types in Freebase, (**?**) builds POS tagging data for low-resource languages by translating and projecting English POS tagging results. (Mintz et al., 2009) builds relation extraction data by considering the sentences containing both the subject and the object of a triple in knowledge base to support the existence of the triple.

Although distant supervision significantly reduces human efforts in build training data, it also introduces noise to the dataset. In relation extraction, (Takamatsu et al., 2012) considers the de-

noising problem as a pre-processing problem, and removes potential noisy sentences by identifying bad syntactic patterns. (Riedel et al., 2010) proposes to alleviate the noise problem by considering the relation extraction task as a multi-instance classification problem based on the assumption that at least one of the retrieved sentences (sentence bag) support the triple. Under the multi-instance classification paradigm, (Hoffmann et al., 2011; Surdeanu et al., 2012) use graphic models to solve the problem, (Zeng et al., 2015) use piece-wise convolutional neural network (PCNN) for sentence embedding and (Lin et al., 2016) further proposes to use attention mechanism to better distinguish true positive ones from false positive ones. Instead of modeling the noise, these models actually only tries to identify the reliable sentences and reduce the influence of unreliable ones.

The at-least-one assumption is not perfect either, and there are still chances that the sentence bag may be false positive or false negative. (Xu et al., 2013) uses pseudo-relevance feedback to find possible false negative data. (Ritter et al., 2013; Min et al., 2013) model the noise by adding a set of latent variables to the MultiR model (Hoffmann et al., 2011) and the MIML model (Surdeanu et al., 2012) respectively to represent the true relation before the observed relations. Our method shares similar spirit to (Ritter et al., 2013) and (Min et al., 2013) in that we all generate the true relation before the observed relation. We differs from them in that our model is designed in the neural network framework and their models are designed in the probabilistic graphic model framework. Furthermore, our model is fully differentiable can model fine grained transition from true relation to observed relation, while their noise modeling parameters need to by set by hand or heuristics and their methods only handle false negative and false positive noise.

Our method is also related to the thread of work of designing denoising neural network component in computer vision. (Sukhbaatar et al., 2014) proposes to use a global transition matrix to transform the true label distribution to the observed label distribution and uses weight decay on the transition matrix during training. (Reed et al., 2014) also uses a hidden layer to represent the true label distribution but try to force it to predict both the noisy label and the input. (Chen and Gupta, 2015; Xiao et al., 2015) first estimates the transition matrix on the clean data set and use it in the noisy data set. (Misra et al., 2016) generates the transition matrix dynamically for each training instance.

In natural language processing (NLP), the research on denoising with neural network is limited. (Fang and Cohn, 2016) also uses a global transition matrix to model the noise introduced by cross-lingual projection of training data in the task of POS tagging for low-resource languages and train the basic model on the clean data first and add the transition matrix when using noisy data afterwards. Our model shares similar spirit with (Misra et al., 2016) in that we all dynamically generate a transition matrix for each training data. However, instead of using EM, we train our model with curriculum learning and trace regularization. Furthermore, we also discuss the architecture for generating bag level transition matrix. Similar to (Fang and Cohn, 2016), we also have both reliable and unreliable data in one of our datasets and we also train on the reliable subset first and then add the unreliable subset. We differs from them in that we also utilize trace regularization to help the transition matrix model the noise better.

8 Conclusion

TODO: re-write In this paper, we propose that the input data may contain useful clues that indicate the noise pattern introduced by distant supervision. We therefore propose to model the noise by dynamically generating a transition matrix for each training data. To overcome the lack direct guidance over the noise pattern, we propose to use curriculum learning to model the noise gradually and trace regularization to help control the behavior of the transition matrix. The experiments on two datasets show that our transition matrix method improves both the sentence level and bag level model in the situation with and without prior knowledge of data quality. We also find that the prior knowledge of data quality can help the training of the transition matrix.

References

Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 1431–1439.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. *arXiv preprint arXiv:1607.01133* .

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1(12).

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*. volume 1, pages 2124–2133.

Bingfeng Luo, Yansong Feng, and Dongyan Zhao. 2016. Improving first order temporal fact extraction with unreliable data. In *Natural Language Processing and Chinese Computing*. Springer.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *HLT-NAACL*. pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels .

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596* .

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 148–163.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1524–1534.

Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics* 1:367–378.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080* .

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. 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*. Association for Computational Linguistics, pages 455–465.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 2691–2699.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *ACL (2)*. pages 665–670.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. EMNLP.