Uncertainty-Aware Self-Training for Semi-Supervised Event Temporal Relation Extraction

Pengfei Cao[1,2], Xinyu Zuo[1,3], Yubo Chen[1,2], Kang Liu[1,2], Jun Zhao[1,2], Wei Bi[3]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, [3]Tencent
{pengfei.cao,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn,{xylonzuo,victoriabi}@tencent.com

ABSTRACT

Extracting event temporal relations is an important task for natural language understanding. Many works have been proposed for supervised event temporal relation extraction, which typically requires a large amount of human-annotated data for model training. However, the data annotation for this task is very time-consuming and challenging. To this end, we study the problem of *semi-supervised event temporal relation extraction*. Self-training as a widely used semi-supervised learning method can be utilized for this problem. However, it suffers from the noisy pseudo-labeling problem. In this paper, we propose the use of uncertainty-aware self-training framework (UAST) to quantify the model uncertainty for coping with pseudo-labeling errors. Specifically, UAST utilizes (1) *Uncertainty Estimation* module to compute the model uncertainty for pseudo-labeling unlabeled data; (2) *Sample Selection with Exploration* module to select informative samples based on uncertainty estimates; and (3) *Uncertainty-Aware Learning* module to explicitly incorporate the model uncertainty into the self-training process. Experimental results indicate that our approach significantly outperforms previous state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**;

KEYWORDS

Event Temporal Relation Extraction, Self-Training, Uncertainty

ACM Reference Format:
Pengfei Cao[1,2], Xinyu Zuo[1,3], Yubo Chen[1,2], Kang Liu[1,2], Jun Zhao[1,2], Wei Bi[3]. 2021. Uncertainty-Aware Self-Training for Semi-Supervised Event Temporal Relation Extraction. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482207

1 INTRODUCTION

Event temporal relation extraction (ETRE) aims to identify temporal relations among mentioned events within a given text. Figure 1 provides a representative example of this task where an ETRE model should be able to predict all pairwise relations among mentioned events, i.e., *explosion* is *BEFORE investigating*, *explosion* is *INCLUDED* in *attack*, and *investigating* is *AFTER attack*. The extraction of temporal relations among events is an important natural language understanding (NLU) task and can facilitate a wide range of downstream applications, likely question answering [16], narrative prediction [3], timeline construction [6] and so on.

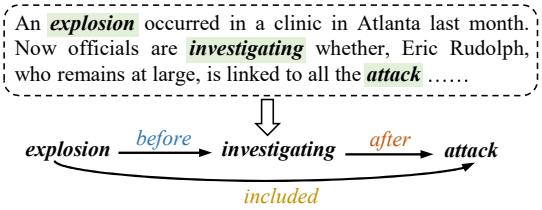

Figure 1: An example of event temporal relation extraction.

In recent years, various neural network models have been proposed for supervised event temporal relation extraction [9, 20, 24, 33], which heavily relies on *abundant human-annotated data* to yield state-of-the-art results. However, the data annotation for the task is known to be very time-consuming and difficult even for experts, because it needs to understand each event's start and end times within a complicated context [2, 24]. As a result, existing event temporal relation extraction datasets are usually small. For example, the widely used dataset MATRES [25] only contains 183 documents for training, which is far from enough to train large neural network models [11]. To this end, we study the problem of *semi-supervised event temporal relation extraction*, which seeks to leverage a limited number of labeled data and a large amount of unlabeled data for model training.

As a widely used semi-supervised learning method, self-training [28, 30] has recently been shown to obtain state-of-the-art performance for various tasks, including neural machine translation [12], text classification [21] and machine reading comprehension [27]. Its basic idea is to first train a model on some amount of labeled data, and then use the updated model to pseudo-annotate unlabeled data. The original labeled data is augmented with the pseudo-labeled data to re-train the model. The iteration training process is repeated until convergence. Due to its simplicity and effectiveness, we apply the self-training method to the semi-supervised event temporal relation extraction task. However, we find that the *noisy pseudo-labeling* constitutes a critical challenge. Specifically, the pseudo-labeled data is inevitably noisy, while the self-training method leverages all pseudo-labeled data without distinction. As a consequence, the model gradually overfits noisy pseudo-labeled samples, which hinders the performance [18, 34]. In this scenario,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482207

it is natural to perform sample selection, which can enable self-training to better cope with pseudo-labeling errors.

Intuitively, when selecting pseudo-annotated samples, if we only focus on the samples that the model already predicts with high confidence, there is little to gain with self-training, because these samples are very familiar to the model. On the other hand, the samples with less predictive confidence are not reliable for the model, because these samples could be very noisy or too difficult to learn from them. Therefore, the model could benefit from judiciously selecting samples that the model is *uncertain* about. However, it is non-trivial to generate uncertainty estimates for non-probabilistic models like neural networks. Fortunately, we can leverage recent advances in Bayesian deep learning [15] to calculate the model uncertainty for pseudo-labeling.

In this paper, we propose the use of uncertainty-aware self-training (UAST) [22] to quantify the model uncertainty for tackling noisy pseudo-labeling problem. The framework consists of three major components: (1) *Uncertainty Estimation*: we leverage Monte Carlo Dropout [7] to compute the model uncertainty in terms of the expectation and variance of predictive probability; (2) *Sample Selection with Exploration*: based on the uncertainty estimate, we employ two entropy-based strategies to select samples that the model is more or less uncertain about for self-training; (3) *Uncertainty-Aware Learning*: when re-training the model on the selected data, we explicitly consider the variance of each sample to further reduce the impact of pseudo-labeling errors. Experimental results on two widely used datasets indicate that our approach substantially outperforms previous state-of-the-art methods.

2 METHODOLOGY

2.1 Base Model

Following recent works [9, 10, 33], we adopt the pre-trained language model (PLM) based architecture as the base model for event temporal relation extraction. Specifically, given an instance x (i.e., an event pair and its context), we first use a PLM, such as BERT [5] or RoBERTa [19], to generate the contextualized embedding for each token. Then, the token embeddings are further fed into a bidirectional long short-term memory network (BiLSTM) [13]. The outputs of the BiLSTM layer for the two events are concatenated as the contextualized features. Finally, we feed the features into a softmax classifier to predict the temporal relation. We refer to the base model as PLM+BiLSTM. The feature extractor (i.e., PLM and BiLSTM) is denoted as f^W, where W denotes model parameters. Figure 2 shows an overview of the UAST framework. We will illustrate each component in detail.

2.2 Uncertainty Estimation

Suppose $D_l = \{x_i, y_i\}_{i=1}^M$ is a set of M labeled instances, where y_i is the label for the instance x_i. Also suppose $D_u = \{x_j\}_{j=1}^N$ is a set of N unlabeled instances. Given the base model PLM+BiLSTM, we first train it on the labeled data D_l. Then, the updated model is used to predict the label of the unlabeled data D_u. At the same time, we leverage the Monte Carlo Dropout [7] to estimate the model uncertainty for pseudo-labeling.

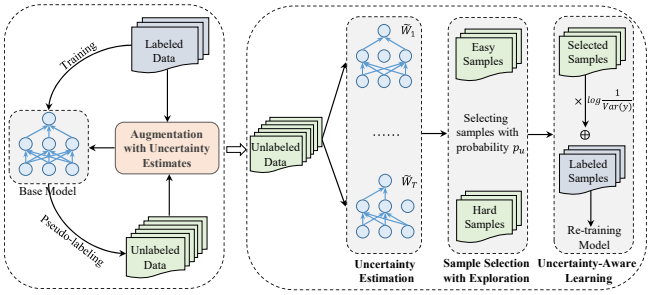

Figure 2: The left part is the overview of the UAST framework. The right part illustrates the detailed procedures that augment the self-training with uncertainty estimates.

Specifically, for each unlabeled instance x_u, we conduct T forward passes with dropout layers being activated. Each pass t generates a pseudo-label denoted as $p({y_t}^*) = \text{softmax}(f^{\widetilde{W}_t}(x_u))$, with corresponding model parameters $\widetilde{W}_t$. We aggregate predictions from T passes to obtain the final pseudo-labels:

$$y = \text{argmax}_c \sum_{t=1}^{T} \mathbb{I}[\text{argmax}_{c'}(p({y_t}^* = c')) = c], \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function. Intuitively, when pseudo-labeling the unlabeled data, the more confident the model is, the higher expectation and lower variance of predictive probability are. Therefore, we consider the model uncertainty in terms of the expectation and variance. Given the results of T forward passes $\{p({y_t}^*)\}_{t=1}^T$, the variance can be approximated by

$$Var(y) = Var[\mathbb{E}(y|x_u; W)] + \mathbb{E}[Var(y|x_u; W)] \approx (\frac{1}{T}\sum_{t=1}^{T} p({y_t}^*)^2 - \mathbb{E}(y)^2), \tag{2}$$

where $\mathbb{E}(y)$ denotes the predictive expectation which can be approximately computed as:

$$\mathbb{E}(y) \approx \frac{1}{T}\sum_{t=1}^{T} p({y_t}^*) = \frac{1}{T}\sum_{t=1}^{T} \text{softmax}(f^{\widetilde{W}_t}(x_u)). \tag{3}$$

2.3 Sample Selection with Exploration

Consider $D'_u = \{x_u, y_u\}$ to be pseudo-labeled dataset. To select informative pseudo-labeled samples, we adopt the Bayesian Active Learning by Disagreement (BALD) measure [14]. The objective of the BALD measure is to select samples that maximize the information gain about the model parameters:

$$\mathbb{B}(y_u, W|x_u, D'_u) = \mathbb{H}[y_u|x_u, D'_u] - \mathbb{E}_{p(W|D'_u)}[\mathbb{H}[y_u|x_u; W]], \tag{4}$$

where $\mathbb{H}[y_u|x_u; W]$ denotes the entropy of predicting the label y_u for sample x_u under the model parameters W. Since the model posterior is intractable, Gal et al. [8] utilize stochastic dropouts to approximate the above measure:

$$\begin{aligned}\mathbb{B}(y_u, W|x_u, D'_u) \approx \widehat{\mathbb{B}}(y_u, W|x_u, D'_u) = &-\sum_c (\frac{1}{T}\sum_t \hat{p}_c^t)\log(\frac{1}{T}\sum_t \hat{p}_c^t) \\ &+ \frac{1}{T}\sum_{t,c} \hat{p}_c^t \log(\hat{p}_c^t),\end{aligned} \tag{5}$$

where $\hat{p}_c^t = p(y_u = c|x_u; \widetilde{W}_t)$. The value of the measure is inversely proportional to the expectation, therefore, a high value of $\widehat{\mathbb{B}}(y_u, W|x_u, D'_u)$ indicates that the model is highly uncertain about

Table 1: Performance comparison with various amounts of labeled data and 50% unlabeled data on the TB-Dense and MATRES.

Methods	TB-Dense						MATRES					
	30%			40%			30%			40%		
	Pre. (%)	Rec. (%)	F1. (%)	Pre. (%)	Rec. (%)	F1. (%)	Pre. (%)	Rec. (%)	F1. (%)	Pre. (%)	Rec. (%)	F1. (%)
PLM+BiLSTM	53.3	53.3	53.3	55.2	55.2	55.2	73.9	69.3	71.5	71.1	74.0	72.5
Han et al. [9]	53.6	53.6	53.6	55.7	55.7	55.7	73.6	70.4	72.0	72.5	73.3	72.9
Mean-Teacher [31]	53.9	53.9	53.9	56.1	56.1	56.1	72.7	71.1	71.9	71.9	74.1	73.0
Self-Training [30]	54.3	54.3	54.3	56.4	56.4	56.4	67.1	77.9	72.1	69.5	77.6	73.3
UAST (Ours)	**58.2**	**58.2**	**58.2**	**60.8**	**60.8**	**60.8**	**74.4**	**76.2**	**75.3**	**73.7**	**79.1**	**76.3**

the predicted label y_u for the sample x_u. Based on the measure, we can employ two strategies to select samples:

(1) ***Selecting Easy Samples:*** We rank the pseudo-labeled samples by $1 - \widehat{\mathbb{B}}(y_u, W|x_u, D'_u)$. The top samples are easier examples, namely, the model is less uncertain about these samples. Intuitively, if we always select these easy samples, the model will not acquire any additional information, because the model is always certain about these examples. Therefore, we select samples with some *exploration* (i.e., probability):

$$p_u = \frac{1 - \widehat{\mathbb{B}}(y_u, W|x_u, D'_u)}{\sum_{x_u \in D'_u} 1 - \widehat{\mathbb{B}}(y_u, W|x_u, D'_u)}. \tag{6}$$

That is to say, we select the instance (x_u, y_u) with probability p_u.

(2) ***Selecting Hard Samples:*** Similar to selecting easy samples, we rank the pseudo-labeled samples by $\widehat{\mathbb{B}}(y_u, W|x_u, D'_u)$. The top samples are called harder ones that the model is more uncertain about. If the model always focuses on these hard samples, it will hinder the performance due to noisy pseudo-labels. Thus, we also select samples with following probability:

$$p_u = \frac{\widehat{\mathbb{B}}(y_u, W|x_u, D'_u)}{\sum_{x_u \in D'_u} \widehat{\mathbb{B}}(y_u, W|x_u, D'_u)}. \tag{7}$$

The above two strategies bias the sampling process towards picking easier samples (i.e., less uncertainty) and harder ones (i.e., more uncertainty), respectively. Our method uses either of the two strategies for selecting samples for self-training.

2.4 Uncertainty-Aware Learning

The above sampling strategies select informative samples according to the posterior entropy. However, the strategies only leverage the expectation, ignoring the predictive variance. To enable the model to better cope with pseudo-labeling errors, we intend to explicitly incorporate the predictive variance into the training process. The original objective of self-training can be formulated as:

$$\min_W \mathbb{E}_{x_l, y_l \in D_l} [-\log p(y_l|x_l; W)] + \mathbb{E}_{x_u \in D_u} \mathbb{E}_{y \sim p(y|x_u; W^*)} [-\log p(y|x_u; W)], \tag{8}$$

where W^* denotes the model parameters after training on the labeled data. In order to incorporate the uncertainty into the self-training, we modify the above loss function as:

$$\min_W \mathbb{E}_{x_l, y_l \in D_l} [-\log p(y_l|x_l, W)] + \mathbb{E}_{x_u \in S_u} \mathbb{E}_{\widetilde{W} \sim q_\theta(W^*)} \mathbb{E}_{y \sim p(y|x_u, \widetilde{W})} \left[-\log p(y|x_u, W) \cdot \log \frac{1}{Var(y)} \right]. \tag{9}$$

where S_u denotes the selected instances. $q_\theta(W^*)$ is the parameter distribution. The model parameters $\widetilde{W} \sim q_\theta(W^*)$ are obtained by activating Dropouts. The per-sample loss for the unlabeled instance x_u is a combination of the log loss $-\log p(y)$ and inverse of its predictive variance given by $\log \frac{1}{Var(y)}$ with log transformation for scaling. We repeat the above procedures (i.e., pseudo-labeling with uncertainty estimation, sample selection and uncertainty-aware learning) until the model convergence.

3 EXPERIMENTS

3.1 Datasets

We evaluate our method on two widely used datasets: (1) **TB-Dense** [1] is constructed based on TimeBank Corpus [29] but addresses the sparse annotation issue in the original data by introducing the *VAGUE* label. It defines 6 classes of temporal relations: *BEFORE*, *AFTER*, *INCLUDES*, *INCLUDED*, *SIMULTANEOUS*, and *VAGUE*. (2) **MATRES** [25] is developed from TempEval-3 [32]. It uses a multi-axis annotation scheme to enhance data quality and adopts a start-point of events to improve inter-annotator agreements. The dataset defines 4 classes of relations: *BEFORE*, *AFTER*, *SIMULTANEOUS*, and *VAGUE*, where the *VAGUE* is regarded as the negative class.

3.2 Evaluation Metrics and Hyperparameters

To be consistent with previous work [23, 24], we adopt micro-averaged precision, recall and F1 score as evaluation metrics. For the TB-Dense, *VAGUE* pairs are taken into consideration (i.e., all classes are seen as positive classes). Thus, the metric should share the same precision, recall and F1 score. For the MATRES, since *VAGUE* pairs are excluded in metrics calculations, the precision, recall and F1 score are different. We leverage the RoBERTa [19] to encode the text. The sizes of hidden states of BiLSTM are 60 and 40 for the TB-Dense and MATRES, respectively. The learning rate is initialized as 2e-5 with a linear decay. We use the Adam algorithm [17] to optimize model parameters.

3.3 Overall Results

Utilizing Limited Labeled Data. For the two datasets, we sample 30% and 40% training data as labeled sets, and we also sample 50% training data as an unlabeled set. We compare our method with two representative semi-supervised learning methods, i.e., Mean-Teacher [31] and Self-Training [30]. In addition, the state-of-the-art ETRE method [9] is also employed as the baseline. Table 1 shows the experimental results on the TB-Dense and MATRES. Overall, we can observe that our method significantly outperforms all the baselines. It demonstrates that our method can effectively alleviate

Table 2: Experimental results on the TB-Dense dataset.

Methods	Pre. (%)	Rec. (%)	F1. (%)
PLM+BiLSTM	62.4	62.4	62.4
Chambers et al. [2]	49.4	49.4	49.4
Cheng and Miyao [4]	52.9	52.9	52.9
Meng and Rumshisky [20]	57.0	57.0	57.0
Han et al. [9]	63.2	63.2	63.2
UAST (Ours)	**64.3**	**64.3**	**64.3**

Table 3: Experimental results on the MATRES dataset.

Methods	Pre. (%)	Rec. (%)	F1. (%)
PLM+BiLSTM	72.1	83.6	77.4
Ning et al. [26]	61.6	72.5	66.6
Ning et al. [25]	66.0	72.3	69.0
Ning et al. [24]	71.3	82.1	76.3
Wang et al. [33]	74.3	**85.0**	78.8
UAST (Ours)	**76.6**	84.9	**80.5**

Table 4: The performance of different sample selection methods on the TB-Dense and MATRES datasets.

Methods	TB-Dense		MATRES	
	30% labeled	40% labeled	30% labeled	40% labeled
Random	54.1	55.9	72.5	73.2
Probability	54.3	55.8	73.1	74.0
Hard (Ours)	56.2	58.4	74.0	74.7
Easy (Ours)	**56.5**	**59.2**	**74.4**	**75.2**

the noisy pseudo-labeling problem of the self-training method.

Utilizing All the Labeled Data. We also evaluate our method on the full training sets. When the model is trained on the TB-Dense dataset, the training set of the MATRES dataset is used as the unlabeled set, vice versa. We compare our method with previous state-of-the-art ETRE models. The results are listed in Table 2 and Table 3. From the results, we can find that our method outperforms all the baselines and achieves state-of-the-art performance on the two datasets. This indicates that leveraging more high-quality labeled data is a direct approach to boost the performance and our method can effectively utilize the unlabeled data for the task.

3.4 Effectiveness of Sample Selection

We compare the effect of different sample selection methods when utilizing 30% and 40% labeled data for the task. The results (i.e., F1 score) are shown in Table 4. "Random" denotes randomly selecting samples from pseudo-labeled data. "Probability" denotes selecting the top-scoring samples based on predictive probabilities. "Easy" and "Hard" denote selecting samples with the probability computed by Eq.(6) and Eq.(7), respectively. From the results, we can observe that the two sample selection strategies both substantially outperform other selection strategies.

Table 5: The performance of different sample weight assignment methods on the TB-Dense and MATRES datasets.

Methods	TB-Dense		MATRES	
	30% labeled	40% labeled	30% labeled	40% labeled
Mean	56.3	59.1	74.1	75.1
Probability	56.5	59.3	74.3	75.2
Uncertainty (Ours)	**58.2**	**60.8**	**75.3**	**76.3**

Table 6: Performance of our method UAST and Self-Training with various amounts of unlabeled data.

Ratio	UAST (Ours)		Self-Training	
	TB-Dense	MATRES	TB-Dense	MATRES
30%	59.5	74.9	56.1	72.3
40%	60.3	75.9	56.3	72.7
50%	60.8	76.3	56.4	73.3
60%	**61.0**	**76.6**	**56.9**	**73.5**

3.5 Effectiveness of Uncertainty Learning

To verify the effectiveness of the uncertainty-aware learning, we compare different sample weight assignment methods. The results are shown in Table 5. "Mean" denotes that all pseudo-labeled instances share same weights. "Probability" denotes that the corresponding probability of predicted label is used as the weight of the sample. From the table, we can observe that the uncertainty-aware learning module achieves the best performance. It indicates the module can further alleviate the impact of noisy pseudo-labeling.

3.6 Varying the Amounts of Unlabeled Data

For semi-supervised ETRE task, we may wonder whether more unlabeled data will always help. To investigate the problem, we fix the amount of labeled data (i.e., 40%) and compare the performance under different amounts of unlabeled data. The results are listed in Table 6. Overall, we can observe that our method and the Self-Training both benefit from a larger amount of unlabeled data. In addition, even utilizing less unlabeled data, our method still achieves better performance than the Self-Training, which indicates that our method can effectively alleviate the noisy pseudo-labeling problem.

4 CONCLUSION

In this paper, we introduce a semi-supervised event temporal relation extraction task and leverage UAST framework for the task. Our method uses the sample selection with exploration module to select informative samples based on the model uncertainty, and utilizes the uncertainty-aware learning module to emphasize low variance samples for self-training. Experimental results demonstrate that our method substantially outperforms previous state-of-the-art models.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (No.2020AAA0106400), National Natural Science Foundation of China (No.61922085), CCF-Tencent Open Research Fund and Beijing Academy of Artificial Intelligence (BAAI2019QN0301).

REFERENCES

[1] Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An Annotation Framework for Dense Event Ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 501–506.

[2] Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense Event Ordering with a Multi-Pass Architecture. *Transactions of the Association for Computational Linguistics* (2014), 273–284.

[3] Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story Comprehension for Predicting What Happens Next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1603–1614.

[4] Fei Cheng and Yusuke Miyao. 2017. Classifying Temporal Relations by Bidirectional LSTM over Dependency Paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 1–6.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[6] Quang Do, Wei Lu, and Dan Roth. 2012. Joint Inference for Event Timeline Construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 677–687.

[7] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, Maria-Florina Balcan and Kilian Q. Weinberger (Eds.). 1050–1059.

[8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Doina Precup and Yee Whye Teh (Eds.). 1183–1192.

[9] Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep Structured Neural Network for Event Temporal Relation Extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 666–106.

[10] Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 434–444.

[11] Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain Knowledge Empowered Structured Neural Net for End-to-End Event Temporal Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5717–5729.

[12] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting Self-Training for Neural Sequence Generation. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997), 1735–1780.

[14] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).

[15] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5574–5584.

[16] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question Answering as Global Reasoning Over Semantic Abstractions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). 1905–1914.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, Yoshua Bengio and Yann LeCun (Eds.).

[18] Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning Dual Retrieval Module for Semi-supervised Relation Extraction. In *The World Wide Web Conference, WWW 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). 1073–1083.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Yuanliang Meng and Anna Rumshisky. 2018. Context-Aware Neural Model for Temporal Information Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 527–536.

[21] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9006–9017.

[22] Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems* (2020).

[23] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint Reasoning for Temporal and Causal Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2278–2288.

[24] Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An Improved Neural Baseline for Temporal Relation Extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6203–6209.

[25] Qiang Ning, Hao Wu, and Dan Roth. 2018. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1318–1328.

[26] Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A Tool for Understanding Time in Natural Language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 72–77.

[27] Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3916–3927.

[28] Gerhard Paass. 1993. Assessing and improving neural network predictions by the bootstrap algorithm. In *Advances in Neural Information Processing Systems*. 196–203.

[29] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. [n. d.]. The timebank corpus.

[30] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-Supervised Self-Training of Object Detection Models. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*. 29–36.

[31] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 1195–1204.

[32] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 1–9.

[33] Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint Constrained Learning for Event-Event Relation Extraction. In *Proceedings of EMNLP*. 696–706.

[34] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.