AdaReNet: Adaptive Reweighted Semi-supervised Active Learning to Accelerate Label Acquisition

Ismini Lourentzou
ilourentzou@vt.edu
Virginia Tech
Blacksburg, VA, 24060, USA

Daniel Gruhl
dgruhl@us.ibm.com
IBM Research Almaden
San Jose, CA, 95120, USA

Alfredo Alba
aalba@us.ibm.com
IBM Research Almaden
San Jose, CA, 95120, USA

Anna Lisa Gentile
annalisa.gentile@ibm.com
IBM Research Almaden
San Jose, CA, 95120, USA

Petar Ristoski
pristoski@ebay.com
eBay Inc
San Jose, CA, 95120, USA

Chad Deluca
delucac@us.ibm.com
IBM Research Almaden
San Jose, CA, 95120, USA

Steven R. Welch
welchs@us.ibm.com
IBM Research Almaden
San Jose, CA, 95120, USA

ChengXiang Zhai
czhai@illinois.edu
University of Illinois at
Urbana-Champaign
Urbana, IL, 61801, USA

ABSTRACT

Data scarcity and quality pose significant challenges to supervised learning. The process of generating informative annotations can be time-consuming and often requires high domain expertise. Active and semi-supervised learning methods can reduce labeling effort by either automatically expanding the training set or by selecting the most informative examples to request domain expert annotation. As most selection methods are heuristic, the performance varies widely across datasets and tasks. Bootstrapping approaches such as self-training can result in negative effects due to the addition of incorrectly pseudo-labeled instances. In this work, we take a holistic approach to label acquisition and consider the expansion of clean and pseudo-labeled subsets jointly. To address the challenge of producing high-quality pseudo-labels, we introduce a collaborative teacher-student framework, where the teacher, termed AdaReNet, learns a data-driven curriculum. Experimental results on several natural language processing (NLP) tasks demonstrate that the proposed framework outperforms baselines.

CCS CONCEPTS

• **Computing methodologies** → **Active learning settings**; **Semi-supervised learning settings**; **Information extraction**.

KEYWORDS

Semi-supervised Learning, Active learning, Neural Networks, Curriculum Learning, Self-training, Pseudo-labeling, Information Extraction, Sequence Labeling

ACM Reference Format:
Ismini Lourentzou, Daniel Gruhl, Alfredo Alba, Anna Lisa Gentile, Petar Ristoski, Chad Deluca, Steven R. Welch, and ChengXiang Zhai. 2021. AdaReNet: Adaptive Reweighted Semi-supervised Active Learning to Accelerate Label Acquisition. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3453892.3461321

1 INTRODUCTION

Deep Learning has been successfully applied in a variety of natural language processing (NLP) tasks, from dependency parsing [58] and named entity recognition [31] to semantic role labeling [19], *etc.* A crucial component is the availability of annotated data. Obtaining labeled examples that can capture the task characteristics is one of the most important prerequisites for supervised learning. Acquiring labels for a large pool of instances in highly technical domains can quickly become prohibitive due to cost, time and expertise requirements. On the other hand, it is often easier to collect inexpensive lower quality *weak* labels through distant supervision, crowd-sourcing, *etc.*, but research has shown that deep neural networks trained on noisy labeled data tend to overfit [53, 61].

Active and semi-supervised learning methods reduce the dependency on large quantities of labeled data. Active learning minimizes annotation by identifying informative subsets of instances with high training utility. Various acquisition strategies have been proposed, but the literature shows that there is no "one-fits-all" solution, and which strategy is the best depends on the downstream task [38, 39]. Semi-supervised methods address the lack of annotations by leveraging unlabeled examples, either by explicitly creating additional training data or by incorporating regularization terms computed on unlabeled instances. These methods can be easily applied

PETRA 2021, June 29-July 2, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3461321

to a variety of tasks, *e.g.*, medical image classification and segmentation [1, 24, 26], land cover classification [46], community-based question answering[59], production lines [11, 12], transportation planning and traffic management [10], and language understanding and generation [5]. Here, we mostly focus on information extraction tasks. Information extraction (IE) can directly enhance pervasive technologies with significant societal impact, *e.g.*, in low-resource settings, IE can be particularly useful for enhancing conversational agents in smart environments [52].

Self-training is one of the earliest methods that enlarge the training set with pseudo-labels. Self-training accepts the classifier's predictions as correct labels when the model is very confident in its predictions [40]. The major drawback of relying on model confidence is the high percentage of incorrectly labeled instances that might be added to the training data. Especially for a model trained on small sets of annotated data, model confidence might not be indicative of correctness [17].

Recent works utilize unlabeled instances as an additional regularizer that incorporates information about the data "manifold" to produce better decision boundary estimates, typically enforcing consistency between the model predictions for the same instance under different noise variations, surpassing supervised learning methods in computer vision classification tasks [44]. Consistency-based methods have been designed for model robustness, after a small static annotated dataset has been acquired, in addition to the large-scale unlabeled data at hand. Similarly, several works demonstrate improved training with large sets of noisy labels [33, 60], with the best-performing methods relying on small trusted fixed datasets for noise-robust training [20, 23, 47]. However, the question of whether we can achieve the same performance gains in a realistic human-in-the-loop scenario with iterative label acquisition remains fairly unanswered.

When considering data acquisition strategies, it is most valuable, for example, to request annotations on instances in which the model has high classification uncertainty, while relying on pseudo-labels for trivial instances. Due to the practical benefits, there has been significant ongoing research in designing active learning heuristics, robust semi-supervised methods, and combining both [15, 45, 51]. In contrast with prior work, we leverage noise-robust training methods in conjunction with semi-supervised learning when the domain expert gradually builds the training dataset, simulating a realistic scenario of label acquisition.

To this end, we propose an effective method that addresses the problem of noisy pseudo-labels generated by the model, with an auxiliary teacher that provides a data-driven curriculum. The teacher model is trained on the domain expert annotated subset and the student predictions, effectively learning to distinguish between correct and noisy labels. On the other hand, the student generates pseudo-labels based on a curriculum determined by the teacher. Both networks are jointly optimized with stochastic gradient descent on streams of domain expert annotations and pseudo-labeled instances, both expanding the training data in each iteration.

The proposed framework enables an efficient parallelizable combination of active and semi-supervised learning that achieves high accuracy, filters out noisy pseudo-labels, and is agnostic to the underlying strategies used for collecting pseudo-labels and domain expert annotations. The contributions of our work are summarized as follows:

- We design a collaborative student-teacher framework that filters out pseudo-labeled instances with a data-driven curriculum strategy. Unlike previous work in semi-supervised research, where the trusted labeled set is static and predefined, we jointly expand the labeled data along the training process.
- We propose an auxiliary teacher, termed **Ada**ptive **Re**weight **Net**work (AdaReNet), that can be combined with any existing model and can be trained jointly under the same computational pipeline, without additional changes to the underlying model.

We validate the robustness of the proposed AdaReNet under a variety of settings, varying the active learning strategies and benchmark tasks. Experimental results show that the proposed method is effective for filtering noisy-labeled instances, outperforms pseudo-labeling baselines, and produces comparable results with semi-supervised regularization techniques.

2 RELATED WORK

Active Learning aims at incorporating *targeted* human annotations: the learning strategy queries an oracle for annotations of specific data points in an iterative manner, where selection criteria, otherwise termed acquisition functions, identify the best data to annotate next. We refer the reader to a review of the most commonly used acquisition functions [49]. The effectiveness of these criteria is highly dependent on the underlying data; it is often very difficult to identify strong connections between any of the criteria and the task at hand [22, 37].

Semi-supervised Learning encompasses algorithms that utilize small amounts of labeled data together with large sets of unlabeled data [6]. A category of these algorithms can be described as generating pseudo-labels by leveraging the model's prediction, but they often produce labels that can be noisy and have been found damaging for several NLP tasks [7, 9]. Recent methods impose some form of noise by relying on model stochasticity [30, 54], data augmentations or perturbations [3, 41, 57], and are commonly evaluated with predefined fixed small training sets.

The combination of active and semi-supervised learning (**Semi-supervised Active learning**) can be used to avoid annotating instances whose labels can be reliably assigned by the learned classification model [43]. Approaches that augment the training data with pseudo-labels often fail when applied to sequence labeling tasks [45], where large amounts of high-quality training data are required to achieve acceptable performance. Including too many tagging errors prevents learning a high-performant model. A solution is for humans to review and correct machine-labeled examples [45], but this can be as costly and time-consuming as acquiring annotations from scratch. Other works request annotations only for the most informative subsequences and automatically label the rest of the sequence [56]. Humans, however, rely on semantics and context to process linguistic information, thus it is often desirable to present the full sequence rather than smaller parts, to prevent annotation ambiguity.

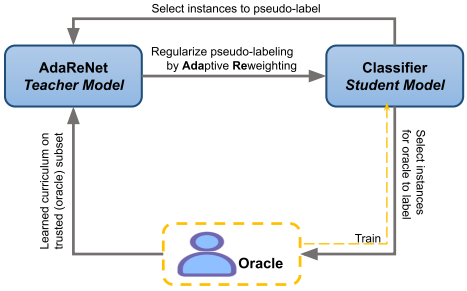


Figure 1: Overall description of the framework. Both the oracle O and the model h_θ produce (pseudo-)labels that are used for training. In addition, oracle annotations are used for training the AdaReNet teacher c_ϕ, that produces a data-driven filtering strategy for the pseudo-labeled data.

Inspired by **Curriculum and Self-paced Learning** [2, 29], recent work learns to re-weight examples [18, 23, 27, 36, 47]. Most methods, however, overlook the annotation process. In realistic scenarios, label acquisition and learning are two interconnected parts of a continuous iterative process. To this end, we analyze the effect of pseudo-labeling when combined with ground-truth collection strategies, and propose a calibrating teac her model to filter out pseudo-labels that are likely to be incorrect.

3 PROBLEM DEFINITION

We are given access to a small set of N labeled instances $D_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ that consists of pairs of input sequences $\mathbf{x} \in \mathcal{X}$, *e.g.*, feature representations of sentences and output labels $\mathbf{y} \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of class labels, *e.g.*, one label per instance in classification tasks or a sequence of T labels $\mathbf{y} = [y_1, y_2, \ldots, y_T]$ in sequence labeling tasks. We denote the total number of unique classes as C. Furthermore, we are given a large pool of unlabeled data $D_U = \{\mathbf{x}_i\}_{i=1}^M$ where M is the number of unlabeled instances and $M \gg N$, an oracle O (domain expert annotator) that we can query for labels and a separate validation dataset D_V. The goal is to learn a model h_θ via utilizing the unlabeled data D_U as much as possible, to produce a more accurate model from what would have been by learning only with oracle annotations. In a sense, the unlabeled set provides useful information about the data "manifold", and improves h_θ's decision boundary estimation. In this work, we utilize unlabeled data in the form of pseudo-labels *alongside the expansion* of the oracle-annotated data, to minimize the labeling effort in every iteration.

Consider a classification task, where the model h_θ is initially trained with a loss $\mathcal{L}_s(D_L, \theta) = \min_\theta \frac{1}{N} \sum_{i=1}^N \mathcal{L}_s(\mathbf{y}_i, h_\theta(\mathbf{x}_i))$, typically cross-entropy for classification tasks, with θ the weights of the neural network. A typical pool-based active learning setting consists of iterations between querying the oracle O for labels on a batch of k instances from the unlabeled data $\{\mathbf{x}_i\}_{i=1}^k$ and retraining the model. With respect to which unlabeled instances to pass for annotation, several active learning acquisition functions have been proposed in the literature [49]. For example, uncertainty sampling selects instances for which the model is the least confident.

We additionally want to minimize the oracle's labeling effort by automatically producing pseudo-labels. As such, in each iteration, the model selects a batch of the m-most *confident* unlabeled examples to assign an inferred pseudo-label, based on the model prediction, with the intention to include these additional data points to the training set and retrain the model [40]. We denote the pseudo-label of an example $\mathbf{x}_i$ as $\hat{\mathbf{y}}_i = \arg\max_j [h_\theta(\mathbf{x}_i)]_j$, where $[\cdot]_j$ corresponds to the class index.

To mitigate potential labeling noise that could diverge training, the distribution of pseudo-labels needs to be consistent with the oracle-generated labels. However, as previously noted, the process of acquiring domain expert labels, in reality, is iterative and baked into the learning process. Here, we incorporate a teacher network c_ϕ to approximate a curriculum, *i.e.*, re-weight the training instances. The teacher takes as input the feature representation of each instance and the student (*i.e.*, classifier) predictions and outputs the learned curriculum, *i.e.*, $w_i = c_\phi(\mathbf{x}_i, h_\theta(\mathbf{x}_i))$. This enables the integration of additional information, for example, student confidence or layer weights; we leave this to future work. The teacher model, termed AdaReNet, is a simple extension that filters out any noisy pseudo-labeled data. Figure 1 illustrates the workflow of the proposed framework. At each training iteration, we optimize the following loss:

$$\min_{\theta \in \mathbb{R}^d, \phi \in \{0,1\}^{M \times C}} \frac{1}{|D_s|} \sum_{i \in D_s} \mathcal{L}_s(\mathbf{y}_i, h_\theta(\mathbf{x}_i)) + \frac{1}{|D_{U'}|} \sum_{i \in D_{U'}} \Big(c_\phi(\mathbf{x}_i, h_\theta(\mathbf{x}_i)) \mathcal{L}_s(\hat{\mathbf{y}}_i, h_\theta(\mathbf{x}_i)) - \lambda c_\phi(\mathbf{x}_i, h_\theta(\mathbf{x}_i)) \Big), \tag{1}$$

where $\mathcal{L}_s$ denote the student (classification) loss function, $D_{U'}$ is the set of pseudo-labeled instances and D_s includes the labeled data from previous iterations, as well as the oracle-labeled data of this iteration. Replacing the teacher c_ϕ with predefined regularization, *e.g.*, ℓ_1-norm, results in the self-paced learning method $c_\phi = \mathbb{1}\{\mathcal{L}_s(\hat{\mathbf{y}}_i, h_\theta(\mathbf{x}_i)) < \lambda\}$, where $\mathbb{1}$ is the indicator function [29]. However, self-paced learning will rely heavily on the proper selection of λ. If λ is too small, then only a few pseudo-labeled instances will be considered. If λ is too large, a large amount of pseudo-labeled data will be added to the training. Recent work that evaluates neural network-based curriculum mechanisms on supervised computer vision tasks has shown that data-driven curricula can choose this hyper-parameter effectively, balancing the trade-off between "easy" and "hard" examples [23]. Other works propose similar mechanisms for mitigating class imbalance [36]. Here, we design AdaReNet to reduce domain expert labeling effort, particularly for sequence labeling tasks. The training framework is summarized in Algorithm 1.

3.1 AdaReNet Teacher

AdaReNet is trained on the 'clean' subset of the training data, *i.e.*, the instances annotated by oracle O, essentially "imitating" the oracle, re-weighting pseudo-labeled data and enforcing consistency with the oracle-generated labels. Intuitively, AdaReNet produces a *latent* confidence weight vector. For each instance $\mathbf{x}_i$ in this 'clean' subset, we can recover both the model prediction $h_\theta(\mathbf{x}_i)$ and the ground-truth label $\mathbf{y}_i$. Thus, we can leverage this information to train

Algorithm 1 Training Algorithm

Input: labeled dataset $D_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, unlabeled pool $D_U = \{\mathbf{x}_i\}_{i=1}^{M}$, Oracle O, Budget B, batch (sample) sizes k, m. QueryAL and QuerySSL denote active and semi-supervised data acquisition functions, respectively.
Output: labeled D_s, trained models h_θ and AdaReNet c_ϕ.

Train student h_θ and AdaReNet c_ϕ on D_L
Initialize $D_s \leftarrow D_L, D_t \leftarrow D_L$
while $|D_s| \leq B$ and $D_U \neq \emptyset$ **do**
  $k = \min\{k, |D_U|\}$ ▷ Adjust if $|D_U| < k$
  $\{\mathbf{x}_i\}_{i=1}^{k} \leftarrow \text{QueryAL}(D_U, k, h_\theta)$ ▷ Select instances to label
  for i=1:k **do**
    $\mathbf{y}_i \leftarrow O(\mathbf{x}_i)$ ▷ Query Oracle for labels
    $D_s \leftarrow D_s \cup \{(\mathbf{x}_i, \mathbf{y}_i)\}$ ▷ Update datasets
    $D_t \leftarrow D_t \cup \{(\mathbf{x}_i, \mathbf{y}_i)\}$
    $D_U \leftarrow D_U \setminus \{\mathbf{x}_i\}$
  $m = \min\{m, |D_U|\}$ ▷ Adjust if $|D_U| < m$
  $\{\mathbf{x}_i\}_{i=1}^{m} \leftarrow \text{QuerySSL}(D_U, m, h_\theta)$ ▷ Select pseudo-labeled
  for i=1:m **do**
    $\hat{\mathbf{y}}_i \leftarrow h_\theta(\mathbf{x}_i)$ ▷ Get predicted label
    $D_s \leftarrow D_s \cup \{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}$ ▷ Update datasets
    $D_U \leftarrow D_U \setminus \{\mathbf{x}_i\}$
  Retrain student h_θ using D_s ▷ Update models
  Retrain AdaReNet c_ϕ using D_t
return final h_θ, c_ϕ, D_s

AdaReNet to predict whether for a particular unlabeled instance $\mathbf{x}_i \in D_U$, the model output is correct or not, *i.e.*, $\mathbb{1}\{\mathbf{y}_i = h_\theta(\mathbf{x}_i)\}$. The teacher loss is defined as $\frac{1}{|D_t|}\sum_{i \in D_t} \mathcal{L}_t(\mathbf{y}_i, c_\phi(\mathbf{x}_i, h_\theta(\mathbf{x}_i)))$, where D_t includes all oracle labeled data thus far.

In terms of model architecture, for sequence labeling tasks, we experiment with two options: 1) $w_i = \sigma(\mathcal{F}(\mathbf{v}^T \mathbf{x}_i \circ h_\theta(\mathbf{x}_i)))$, where $\circ$ is the Hadamard product, $\mathcal{F}$ is a convolutional neural network with multiple filters, $\mathbf{v}^T$ is a feed-forward layer that embeds the input and σ is a final feed-forward layer with a sigmoid activation function and 2) $w_i = \sigma([\mathcal{F}^{in}(\mathbf{x}_i); \mathcal{F}^{pr}(h_\theta(\mathbf{x}_i))])$ where $[;]$ denotes concatenation, $\mathcal{F}^{in}$ and $\mathcal{F}^{pr}$ are convolutional neural networks with multiple filters that embed the instance representation and the student predictions. In our preliminary experiments, we found that the second approach routinely outperforms the first, so we present results accordingly [1]. Additional model details are provided in Section 4.3.

4 EXPERIMENTS

We describe the experimental setup, *e.g.*, model architectures and hyper-parameters. We evaluate AdaReNet across a variety of NLP sequence labeling and classification tasks, models (neural and traditional) and data sizes. In addition, we compare with state-of-the-art semi-supervised methods (combined with active learning) and a pure active learning setting where data are labeled solely by the oracle, *i.e.*, all selected instances in each iteration are labeled by a domain expert. We note that the framework is fairly general and can be applied to other tasks such as speech recognition and energy disaggregation [16, 25].

4.1 Datasets

Named Entity Recognition (NER): We use the CONLL 2003 data set [55] that includes text instances annotated with Person (*PER*), Location (*LOC*), Organization (*ORG*) and Miscellaneous (*MISC*) entities. We keep the original annotation, which is based on IOB1 labeling [2]. The dataset is partitioned into *train*, *testa* and *testb*, with 14987, 3466 and 3684 instances, respectively. We divide *train* into a small labeled set D_L and an unlabeled set D_U (labels remain hidden until they are requested from the oracle). Additionally, we treat *testa* as validation D_V and test the final model (trained on data collected from each experiment) on *testb*.
Part-of-Speech Tagging (POS): We use the CONLL 2003 data set (POS labels) [55], with the same splits as for NER.
Text Chunking (CHUNK): We make use of the CONLL 2000 dataset [48] that contains annotated text from the WSJ corpus. The number of training instances is 8936, and the testing instances are 2012. There is no predefined split; we randomly sample $\approx 10\%$ of the training instances as validation data.
Question classification (QC): We evaluate on the (TREC-6) question classification dataset [35] consisting of open-domain fact-based questions classified into six semantic categories. The dataset contains 5452 training examples and 500 test examples. We use $\approx 10\%$ as validation data.

For NER and CHUNK, the evaluation metric is F1, as defined by CoNLL [55], *i.e.*, only exact matches between actual and predicted entities are counted as correct. For POS and QC, the evaluation metric is accuracy.

4.2 Active Learning Acquisition Functions

The data acquisition functions for collecting oracle annotations will largely influence the AdaReNet model capability of generating useful data-driven curricula. Thus, we experiment with two active learning strategies, covering two main categories: (i) density-based acquisition strategies, where the geometry of the feature space is used for selecting diverse samples [14], and (ii) model-based, where the model prediction is used to calculate uncertainty estimates for instance selection [34].
Diversity sampling (DS): the selected data point is the most diverse of all instances already labeled, *i.e.*, the similarity between the labeled data and the chosen data point is minimized:

$$\arg\min_{\mathbf{x} \in D_U} \sum_{\mathbf{x}' \in D_s} \delta\left(h_\theta(\mathbf{x}), h_\theta(\mathbf{x}')\right) \tag{2}$$

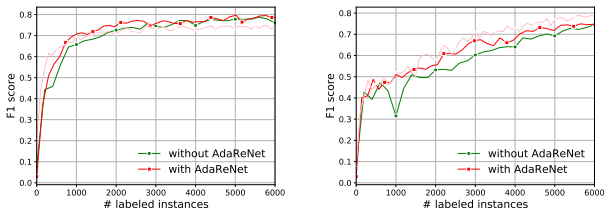
Uncertainty sampling (US): The instances for which the model is the most uncertain are selected. The most commonly used measure of uncertainty is entropy, where the acquisition strategy becomes

$$\arg\max_{\mathbf{x} \in D_U} -\sum_{j=1}^{C} p_\theta(y_j|\mathbf{x}) \log p_\theta(y_j|\mathbf{x}), \tag{3}$$

where C is the number of classes and the model h_θ is used to calculate $p_\theta(\mathbf{y}|\mathbf{x})$. For sequence labeling tasks, we use the Total

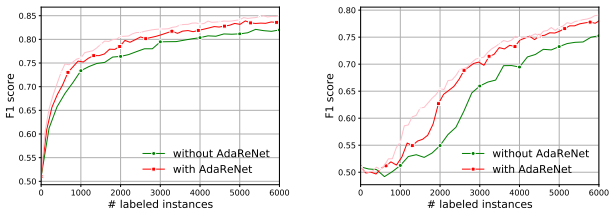
[1] We further experimented with substituting the convolutions with an LSTM layer; there was little variation w.r.t. results, an outcome that is on-par with prior work [23].

[2] https://lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo/



(a) Uncertainty (b) Diversity

Figure 2: Comparison on NER with a charLSTM-biLSTM-CRF student (classifier) model and (red) or without (green) the AdaReNet teacher, when oracle annotations are collected with Uncertainty (left), and Diversity (right). The pink lines represent pure active learning data collection. Best viewed in color.



(a) Uncertainty (b) Diversity

Figure 3: Replacing the NER student (classifier) model with a traditional CRF. Results with (red) or without (green) the AdaReNet teacher, with Uncertainty (left) and Diversity (right). The pink lines represent pure active learning data collection. Best viewed in color.

Token Entropy (TTE) [50]:

$$\arg\max_{\mathbf{x}\in D_U} -\sum_{i=1}^{T}\sum_{j=1}^{C} p_\theta(y_{ij}|x_i)\log p_\theta(y_{ij}|x_i), \tag{4}$$

where T is the total sequence length.

4.3 Implementation Details

For sequence labeling tasks, h_θ is implemented as a charLSTM-biLSTM-CRF model [8]. For the classification tasks, h_θ is a word-level GRU classifier. Models are initialized with 100-dimensional pre-trained Glove embeddings. The charLSTM-biLSTM-CRF model includes 25-dimensional character-level embeddings and additional casing embeddings that map words into common representations based on the occurrence of digits, capitals or lowercase terms [8]. The word-level, character-level and casing emdeddings are concatenated and followed by 0.05 word dropout, 0.5 variational dropout [13] and one biLSTM layer with 256 hidden neurons, plus a final CRF layer [31]. The biLSTM classifier consists of one GRU layer with 256 hidden neurons, 0.5 dropout [21] and a final linear layer. We train the models with Adam [28], 0.01 initial learning rate, 0.5 learning rate decay and 32 batch-size. For the consistency-based semi-supervised methods, we incorporate Gaussian noise on the embedding level as perturbations.

The initial labeled dataset is 100 instances. At each iteration the active learning and pseudo-labeling components query labels for 100 instances each (*i.e.*, in total 200 newly annotated examples per iteration). Then, AdaReNet filters pseudo-labeled instances accordingly. For consistency-based methods, half of the budget is annotated by an oracle and the rest is used as additional unlabeled instances for calculating the unsupervised consistency loss. We chose an equal annotation batch size for both the oracle and the model, as we have observed that allocating more instances for pseudo-labeling results in decreased performance. With respect to the number of instances labeled in each iteration, preliminary experiments showed that the smaller data annotation batch sizes result in more frequent model updates and greater learning efficiency, findings that are on-par with existing literature [4, 38, 42].

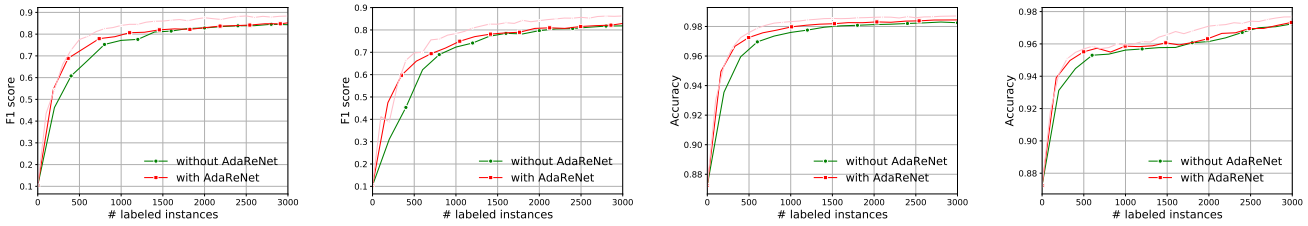
4.4 Experimental Results

Since the initial labeled pool is very small ($\approx 2-3\%$ of the total annotation budget, depending on the task) we anticipate that a lot of incorrectly pseudo-labeled instances may be added to the training. This noise can be detrimental to learning and has been shown to cause overfitting [61]. To this end, we design a teacher model termed AdaReNet, that learns to filter out pseudo-labeled instances with a data-driven curriculum strategy. Our experiments show that the curriculum designed by AdaReNet improves performance.

In Figure, 2 our baseline is the student h_θ (*i.e.*, the charLSTM-biLSTM-CRF model) without the AdaReNet teacher (green lines) when data are collected with uncertainty or diversity sampling. The addition of AdaReNet (red lines) improves model performance. When compared with a *pure active learning scenario*, where all instances are passed to the oracle O to acquire labels (pink lines), the AdaReNet can optimize the trade-off between oracle annotations (domain expert labels) and pseudo-labeling errors to achieve performance closer to the optimal case of collecting all training examples with the oracle. AdaReNet improves performance across all tasks, *e.g.*, including Chunking and Part-of-Speech Tagging (Figure 4). Ultimately, as the pool of oracle annotations grows, the student network becomes better at selecting unlabeled data and at producing pseudo-labels. At this point, removing AdaReNet from the training process is possible. We leave the design of stopping/removal criteria for the teacher to future work.

To reduce the total number of parameters the teacher might add to the training, we designed AdaReNet to share the embedding layer with the student model. To ensure that AdaReNet improves performance beyond what parameter sharing would do, we experiment with the student architecture. More specifically, we replace the student model with a traditional CRF model that shares no weights with the teacher (Figure 3). The performance improvements were amplified. In other words, this result implies that the AdaReNet is effective in learning data-driven curricula that help the student network.

A strong baseline of a predefined curriculum can be based on a strict threshold [32, 40]. For example, we can filter out pseudo-labeled instances that do not surpass a predefined confidence threshold, *e.g.*, $\max_{\mathbf{y}} p_\theta(\mathbf{y}|\mathbf{x}) < 95\%$. This means that only when the class



(a) CHUNK - Uncertainty (b) CHUNK - Diversity (c) POS - Uncertainty (d) POS - Diversity

Figure 4: Comparison on Chunking (CHUNK) and Part-of-Speech Tagging (POS) with (red) or without (green) the AdaReNet teacher, when oracle annotations are collected with active learning (Uncertainty or Diversity). The pink lines represent pure active learning data collection. Best viewed in color.

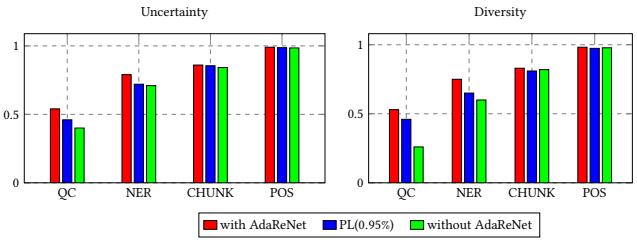


Figure 5: Comparison of the student network with (red) and without the AdaReNet teacher (green), and with a predefined confidence threshold (blue). Oracle (clean) annotations collected with Uncertainty (left) or Diversity (right). x-axis: F1 for {NER,CHUNK}, Accuracy for {POS, QC}. Best viewed in color.

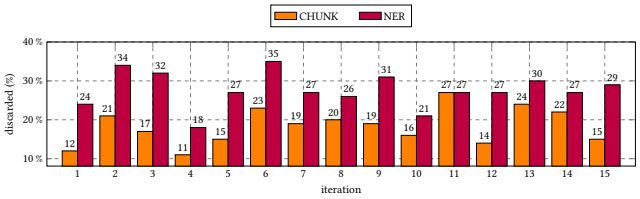


Figure 6: Ratio of discarded incorrect pseudo-labels in each round, on Chunking (CHUNK) and Named Entity Recognition (NER) tasks, with Uncertainty Sampling. Best viewed in color.

Table 1: Comparison with baseline methods.

Method	Uncertainty				Diversity			
	NER	CHUNK	POS	QC	NER	CHUNK	POS	QC
	F1		Acc		F1		Acc	
AdaReNet	**0.79**	0.86	0.99	**0.54**	**0.75**	0.83	0.98	**0.53**
Student	0.71	0.84	0.99	0.40	0.60	0.82	0.98	0.25
PL(0.95%) [32, 40]	0.72	0.86	0.99	0.46	0.65	0.81	0.97	0.46
IImodel [30]	0.68	0.85	0.99	0.46	0.65	0.81	0.98	0.42
ICT [57]	0.72	0.85	0.99	0.41	0.63	0.83	0.98	0.30
MT [54]	0.73	0.86	0.99	0.43	0.64	0.82	0.98	0.44
VAT [41]	0.75	0.85	0.98	0.51	0.63	0.82	0.99	0.41

probability prediction is above that threshold (*i.e.*, the model is almost absolutely certain about its prediction) then the instance is pseudo-labeled and used during training [32, 44]. For sequence labeling, model confidence is defined here as $\left[\max_{\mathbf{y}} p_\theta(\mathbf{y}|\mathbf{x})\right]^{1/T}$ where T is the sequence length. We conduct experiments to evaluate whether such a strict predefined threshold would be sufficient in terms of test performance, *i.e.*, under which conditions there is no need for data-driven curricula. The AdaReNet outperforms such baseline in two tasks, Named Entity Recognition (NER) and Question Classification (QC), but provides relatively small improvements on easier tasks, such as Chunking and Part-of-Speech Tagging (Figure 5 and Table 1). Overall, AdaReNet results in up to 10% performance gains w.r.t. the next best pseudo-labeling strategy.

We also compare with consistency-based semi-supervised methods that incorporate unlabeled data as additional regularization [44]. Given an unlabeled example, Π model [30] computes consistency between two model predictions, under model stochasticity, *e.g.*, dropout. Virtual Adversarial Training (VAT) [41] finds the worst local perturbation that will alter the model predictions the most. Mean Teachers (MT) [54] uses an exponential moving average of model parameters as a teacher that produces the targets for the student model. Finally, Interpolation Consistency Training (ICT) [57] averages over multiple augmented versions of an instance and incorporates MixUp, *i.e.*, linearly interpolating input instances and output labels [62]. AdaReNet surpasses all baselines on {NER, QC} tasks and maintains comparable performance with consistency-based semi-supervised algorithms (Table 1).

Finally, in Figure 6, for each round and two tasks, CHUNK and NER, we present the ratio of noisy pseudo-labeled instances that are discarded versus the remaining ones that the AdaReNet was not able to identify and remove, with the best performing active learning strategy (uncertainty). The AdaReNet teacher exhibits approximately $11-35\%$ labeling noise reduction rates, depending on the task.

5 CONCLUSIONS AND FUTURE WORK

In this work, we present a collaborative teacher-student method that expands the training set with both human and pseudo labels. The teacher, termed AdaReNet, learns a data-driven curriculum strategy to select reliable pseudo-labeled data that can be confidently used during training. In the future, we hope to test our

method with additional semi-supervised methods and batch active learning algorithms that take into account the redundancy among instances.

REFERENCES

[1] Christoph Baur, Shadi Albarqouni, and Nassir Navab. 2017. Semi-supervised deep learning for fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 311–319.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual International Conference on Machine Learning (ICML)*. 41–48.

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*.

[4] Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*. 59–66.

[5] Ernie Chang, Vera Demberg, and Alex Marin. 2021. Jointly Improving Language Understanding and Generation with Quality-Weighted Weak Supervision of Automatic Labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 818–829.

[6] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. 2010. *Semi-Supervised Learning*. The MIT Press.

[7] Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 15th National Conference on Artificial Intelligence and 9th Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI)*. 598–603.

[8] Jason PC Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4 (2016), 357–370.

[9] Stephen Clark, James R Curran, and Miles Osborne. 2003. Bootstrapping POS taggers using unlabelled data. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL - Volume 4*. 49–55.

[10] Sina Dabiri, Chang-Tien Lu, Kevin Heaslip, and Chandan K Reddy. 2019. Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 1010–1023.

[11] Nikolaos Doulamis and Anastasios Doulamis. 2012. Fast and adaptive deep fusion learning for detecting visual objects. In *European Conference on Computer Vision*. Springer, 345–354.

[12] Nikolaos Doulamis and Anastasios Doulamis. 2014. Semi-supervised deep learning for object tracking and classification. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 848–852.

[13] Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)* 29 (2016), 1019–1027.

[14] Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *arXiv:1711.00941* (2017).

[15] Andrew Goldberg, Xiaojin Zhu, Alex Furger, and Jun-Ming Xu. 2011. Oasis: Online active semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 25.

[16] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *IEEE workshop on automatic speech recognition and understanding*. IEEE, 273–278.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.

[18] Jiangfan Han, Ping Luo, and Xiaogang Wang. 2019. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*. 5138–5147.

[19] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 473–483.

[20] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*. 10477–10486.

[21] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[22] Wei-Ning Hsu and Hsuan-Tien Lin. 2015. Active learning by learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[23] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, 2304–2313.

[24] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. 2021. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data* 8, 1 (2021), 1–18.

[25] Maria Kaselimi, Nikolaos Doulamis, Anastasios Doulamis, Athanasios Voulodimos, and Eftychios Protopapadakis. 2019. Bayesian-optimized bidirectional LSTM regression model for non-intrusive load monitoring. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2747–2751.

[26] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, and Nikolaos Doulamis. 2020. Transfer Learning for COVID-19 Pneumonia Detection and Classification in Chest X-ray Images. *medRxiv* (2020).

[27] Tae-Hoon Kim and Jonghyun Choi. 2018. Screenernet: Learning self-paced curriculum for deep neural networks. *arXiv arXiv:1801.00904* (2018).

[28] Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*.

[29] M Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1 (NeurIPS)*. 1189–1197.

[30] Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. In *Proceedings of the 5th International Conference for Learning Representations (ICLR)*.

[31] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 260–270.

[32] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, Vol. 3.

[33] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5447–5456.

[34] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[35] Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan*. 556–562.

[36] Yunru Liu, Tingran Gao, and Haizhao Yang. 2020. Selectnet: Learning to sample from the wild for imbalanced data training. In *Mathematical and Scientific Machine Learning*. PMLR, 193–206.

[37] Ismini Lourentzou, Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, and Steve Welch. 2018. Mining relations from unstructured content. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 363–375.

[38] Ismini Lourentzou, Daniel Gruhl, and Steve Welch. 2018. Exploring the efficiency of batch active learning for human-in-the-loop relation extraction. In *Companion Proceedings of the The Web Conference 2018*. 1131–1138.

[39] David Lowell, Zachary C Lipton, and Byron C Wallace. 2019. Practical Obstacles to Deploying Active Learning. 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)*. 21–30.

[40] David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective Self-Training for Parsing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-HLT)*. 152.

[41] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 1979–1993.

[42] Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. 2014. Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment* 8, 2 (2014), 125–136.

[43] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 1 (2000), 34.

[44] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 31. 3235–3246.

[45] David Pierce and Claire Cardie. 2001. Limitations of Co-Training for Natural Language Learning from Large Datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[46] Eftychios Protopapadakis, Anastasios Doulamis, Nikolaos Doulamis, and Evangelos Maltezos. 2021. Stacked Autoencoders Driven by Semi-Supervised Learning for Building Extraction from near Infrared Remote Sensing Imagery. *Remote Sensing* 13, 3 (2021), 371.

[47] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, 4334–4343.

[48] Erik Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task Chunking. In *Proceedings of the Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

[49] Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* (2012).

[50] Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1070–1079.

[51] Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. 2019. Rethinking deep active learning: Using unlabeled data at model training. *arXiv preprint arXiv:1911.08177* (2019).

[52] Evropi Stefanidi, Maria Korozi, Asterios Leonidis, and Margherita Antona. 2018. Programming intelligent environments in natural language: an extensible interactive approach. In *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference (PETRA)*. 50–57.

[53] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5552–5560.

[54] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*. 1195–1204.

[55] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. 142–147.

[56] Katrin Tomanek and Udo Hahn. 2009. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1039–1047.

[57] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. 2019. Interpolation consistency training for semi-supervised learning. In *Proceedings of the International Joint Conference on Artifical Intelligence (IJCAI)*.

[58] Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2306–2315.

[59] Papis Wongchaisuwat, Diego Klabjan, and Siddhartha Reddy Jonnalagadda. 2016. A semi-supervised learning approach to enhance health care community–based question answering: A case study in alcoholism. *JMIR Medical Informatics* 4, 3 (2016).

[60] 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 (CVPR)*. 2691–2699.

[61] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of the 5th International Conference for Learning Representations (ICLR)*.

[62] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *Proceedings of the 6th International Conference for Learning Representations (ICLR)*.