Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling

Ouyu Lan[†*] **Xiao Huang**[†*] **Bill Yuchen Lin**[†] **He Jiang**[†] **Liyuan Liu**[‡] **Xiang Ren**[†]
{olan,huan183,yuchen.lin,jian567,xiangren}@usc.edu, ll2@illinois.edu
[†]Computer Science Department, University of Southern California
[‡]Computer Science Department, University of Illinois at Urbana-Champaign

Abstract

Sequence labeling is a fundamental task for a range of natural language processing problems. When used in practice, its performance is largely influenced by the annotation quality and quantity, and meanwhile, obtaining ground truth labels is often costly. In many cases, ground truth labels do not exist, but noisy annotations or annotations from different domains are accessible. In this paper, we propose a novel framework *Consensus Network* (CONNET) that can be trained on annotations from multiple sources (e.g., crowd annotation, cross-domain data). It learns individual representation for every source and dynamically aggregates source-specific knowledge by a context-aware attention module. Finally, it leads to a model reflecting the agreement (consensus) among multiple sources. We evaluate the proposed framework in two practical settings of multi-source learning: learning with crowd annotations and unsupervised cross-domain model adaptation. Extensive experimental results show that our model achieves significant improvements over existing methods in both settings. We also demonstrate that the method can apply to various tasks and cope with different encoders. [1]

1 Introduction

Sequence labeling is a general approach encompassing various natural language processing (NLP) tasks including part-of-speech (POS) tagging (Ratnaparkhi, 1996), word segmentation (Low et al., 2005), and named entity recognition (NER) (Nadeau and Sekine, 2007). Typically, existing methods follow the supervised learning paradigm, and require high-quality annotations. While gold standard annotation is expensive and time-consuming, imperfect annotations are relatively easier to obtain from crowdsourcing (noisy labels) or other domains (out-of-domain). Despite their low cost, such supervision usually can be obtained from different sources, and it has been shown that multi-source weak supervision has the potential to perform similar to gold annotations (Ratner et al., 2016).

Specifically, we are interested in two scenarios: 1) **learning with crowd annotations** and 2) **unsupervised cross-domain model adaptation**. Both situations suffer from imperfect annotations, and benefit from multiple sources. Therefore, the key challenge here is to aggregate multi-source imperfect annotations for learning a model without knowing the underlying ground truth label sequences in the target domain.

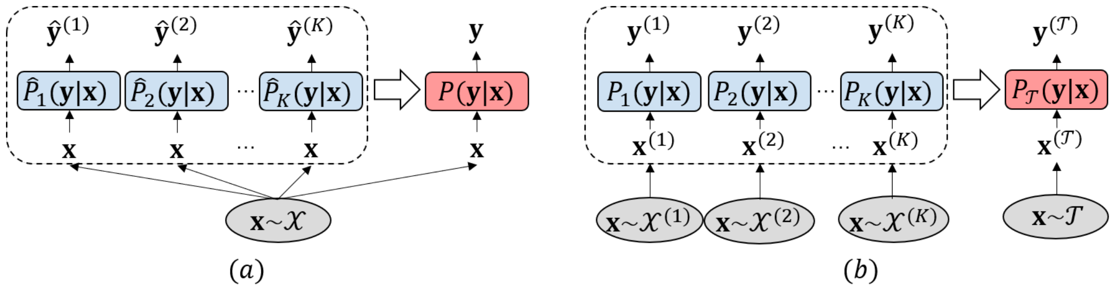
Our intuition mainly comes from the phenomenon that different sources of supervision have different strengths and are more proficient with distinct situations. Therefore they may not keep consistent importance during aggregating supervisions, and aggregating multiple sources for a specific input should be a dynamic process that depends on the sentence context. To better model this nature, we need to (1) explicitly model the unique traits of different sources when training and (2) find best suitable sources for generalizing the learned model on unseen sentences.

In this paper, we propose a novel framework, named *Consensus Network* (CONNET), for sequence labeling with multi-source supervisions. We represent the annotation patterns as different biases of annotators over a shared behavior pattern. Both annotator-invariant patterns and annotator-specific biases are modeled in a decoupled way. The first term comes through sharing part of low-level model parameters in a multi-task learning schema. For learning the biases, we decouple them from the model as the transformations

[*] The first two authors contributed equally.

[1] Our code can be found at `https://github.com/INK-USC/ConNet`.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2134–2146
July 5 - 10, 2020. ©2020 Association for Computational Linguistics



$\hat{\mathbf{y}}^{(1)}$
$\hat{\mathbf{y}}^{(2)}$
$\hat{\mathbf{y}}^{(K)}$
$\mathbf{y}$
$\hat{P}_1(\mathbf{y}|\mathbf{x})$
$\hat{P}_2(\mathbf{y}|\mathbf{x})$
$\hat{P}_K(\mathbf{y}|\mathbf{x})$
$P(\mathbf{y}|\mathbf{x})$
$\mathbf{x}$
$\mathbf{x}\sim\mathcal{X}$
(a)
$\mathbf{y}^{(1)}$
$\mathbf{y}^{(2)}$
$\mathbf{y}^{(K)}$
$\mathbf{y}^{(\mathcal{T})}$
$P_1(\mathbf{y}|\mathbf{x})$
$P_2(\mathbf{y}|\mathbf{x})$
$P_K(\mathbf{y}|\mathbf{x})$
$P_{\mathcal{T}}(\mathbf{y}|\mathbf{x})$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(K)}$
$\mathbf{x}^{(\mathcal{T})}$
$\mathbf{x}\sim\mathcal{X}^{(1)}$
$\mathbf{x}\sim\mathcal{X}^{(2)}$
$\mathbf{x}\sim\mathcal{X}^{(K)}$
$\mathbf{x}\sim\mathcal{T}$
(b)

Figure 1: **Illustration of the task settings for the two applications in this work**: (a) learning consensus model from crowd annotations; (b) unsupervised cross-domain model adaptation.

on top-level tagging model parameters, such that they can capture the unique strength of each annotator. With such decoupled source representations, we further learn an attention network for dynamically assigning the best sources for every unseen sentence through composing a transformation that represents the agreement among sources (consensus). Extensive experimental results in two scenarios show that our model outperforms strong baseline methods, on various tasks and with different encoders. CONNET achieves state-of-the-art performance on real-world crowdsourcing datasets and improves significantly in unsupervised cross-domain adaptation tasks over existing works.

2 Related Work

There exists three threads of related work with this paper, which are sequence labeling, crowdsourcing and unsupervised domain adaptation.

Neural Sequence Labeling. Traditional approaches for sequence labeling usually need significant efforts in feature engineering for graphical models like conditional random fields (CRFs) (Lafferty, 2001). Recent research efforts in neural network models have shown that end-to-end learning like convolutional neural networks (CNNs) (Ma and Hovy, 2016a) or bidirectional long short-term memory (BLSTMs) (Lample et al., 2016) can largely eliminate human-crafted features. BLSTM-CRF models have achieved promising performance (Lample et al., 2016) and are used as our base sequence tagging model in this paper.

Crowd-sourced Annotation. Crowd-sourcing has been demonstrated to be an effective way of fulfilling the label consumption of neural models (Guan et al., 2017; Lin et al., 2019). It collects annotations with lower costs and a higher speed from non-expert contributors but suffers from some degradation in quality. Dawid and Skene (1979) proposes the pioneering work to aggregate crowd annotations to estimate true labels, and Snow et al. (2008) shows its effectiveness with Amazon's Mechanical Turk system. Later works (Dempster et al., 1977; Dredze et al., 2009; Raykar et al., 2010) focus on Expectation-Maximization (EM) algorithms to jointly learn the model and annotator behavior on classification.

Recent research shows the strength of multi-task framework in semi-supervised learning (Lan et al., 2018; Clark et al., 2018), cross-type learning (Wang et al., 2018), and learning with entity triggers (Lin et al., 2020). Nguyen et al. (2017); Rodrigues and Pereira (2018); Simpson et al. (2020) regards crowd annotations as noisy gold labels and constructs crowd components to model annotator-specific bias which were discarded during the inference process. It is worth mentioning that, it has been found even for human curated annotations, there exists certain label noise that hinders the model performance (Wang et al., 2019).

Unsupervised Domain Adaptation. Unsupervised cross-domain adaptation aims to transfer knowledge learned from high-resource domains (source domains) to boost performance on low-resource domains (target domains) of interests such as social media messages (Lin et al., 2017). Different from supervised adaptation (Lin and Lu, 2018), we assume there is no labels at all for target corpora. Saito et al. (2017) and Ruder and Plank (2018) explored bootstrapping with multi-task tri-training approach, which requires unlabeled data from the target domain. The method is developed for one-to-one domain adaptation and does not model the differences among multiple source domains. Yang and Eisenstein (2015) represents each domain with a vector of metadata domain attributes and uses domain vectors to train the model to deal with domain shifting, which is highly dependent on prior domain knowledge.

(Ghifary et al., 2016) uses an auto-encoder method by jointly training a predictor for source labels, and a decoder to reproduce target input with a shared encoder. The decoder acts as a normalizer to force the model to learn shared knowledge between source and target domains. Adversarial penalty can be added to the loss function to make models learn domain-invariant feature only (Fernando et al., 2015; Long et al., 2014; Ming Harry Hsu et al., 2015). However, it does not exploit domain-specific information.

3 Multi-source Supervised Learning

We formulate the multi-source sequence labeling problem as follows. Given K sources of supervision, we regard each source as an imperfect annotator (non-expert human tagger or models trained in related domains). For the k-th source data set $S^{(k)} = \{(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)})\}_{i=1}^{m_k}$, we denote its i-th sentence as $\mathbf{x}_i^{(k)}$ which is a sequence of tokens: $\mathbf{x}_i^{(k)} = (x_{i,1}^{(k)}, \cdots, x_{i,N}^{(k)})$. The tag sequence of the sentence is marked as $\mathbf{y}_i^{(k)} = \{y_{i,j}^{(k)}\}$. We define the sentence set of each annotators as $\mathcal{X}^{(k)} = \{\mathbf{x}_i^{(k)}\}_{i=1}^{m_k}$, and the whole training domain as the union of all sentence sets: $\mathcal{X} = \bigcup_{k=1}^{(K)} \mathcal{X}^{(k)}$. The goal of the multi-source learning task is to use such imperfect annotations to train a model for predicting the tag sequence $\mathbf{y}$ for any sentence $\mathbf{x}$ in a target corpus $\mathcal{T}$. Note that the target corpus $\mathcal{T}$ can either share the same distribution with $\mathcal{X}$ (Application I) or be significantly different (Application II). In the following two subsections, we formulate two typical tasks in this problem as shown in Fig. 1.

Application I: Learning with Crowd Annotations. When learning with crowd-sourced data, we regard each worker as an imperfect annotator ($S^{(k)}$), who may make mistakes or skip sentences in its annotations. Note that in this setting, different annotators tag subsets of the *same* given dataset ($\mathcal{X}$), and thus we assume there are no input distribution shifts among $\mathcal{X}^{(k)}$. Also, we only test sentences in the same domain such that the distribution in target corpus $\mathcal{T}$ is the same as well. That is, the marginal distribution of target corpus $P_{\mathcal{T}}(\mathbf{x})$ is the same with that for each individual source dataset, *i.e.* $P_{\mathcal{T}}(\mathbf{x}) = P_k(\mathbf{x})$. However, due to imperfectness of the annotations in each source, $P_k(\mathbf{y}|\mathbf{x})$ is shifted from the underlying truth $P(\mathbf{y}|\mathbf{x})$ (illustrated in the top-left part of Fig. 1). The multi-source learning objective here is to learn a model $P_{\mathcal{T}}(\mathbf{y}|\mathbf{x})$ for supporting inference on any new sentences in the same domain.

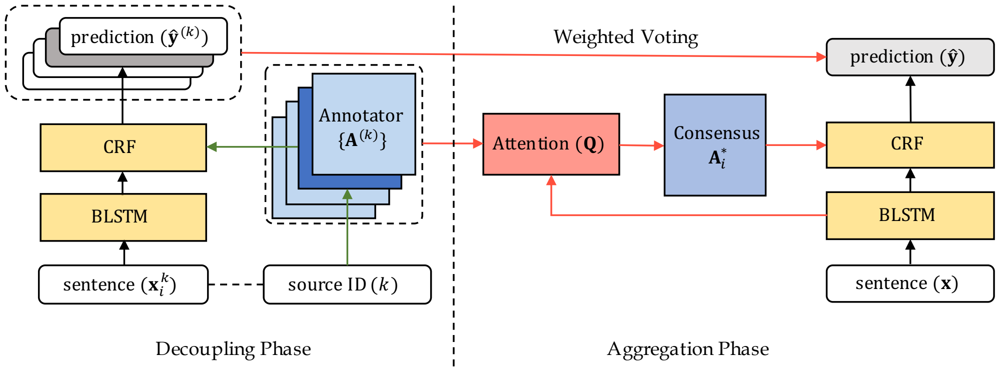
Application II: Unsupervised Cross-Domain Model Adaptation. We assume there are available annotations in several source domains, but not in an unseen target domain. We assume that the input distributions $P(\mathbf{x})$ in different source domains $\mathcal{X}^{(k)}$ vary a lot, and such annotations can hardly be adapted for training a target domain model. That is, the prediction distribution of each domain model ($P_k(\mathbf{y}|\mathbf{x})$) is close to the underlying truth distribution ($P(\mathbf{y}|\mathbf{x})$) only when $\mathbf{x} \in \mathcal{X}^{(k)}$. For target corpus sentences $\mathbf{x} \in \mathcal{T}$, such a source model $P_k(\mathbf{y}|\mathbf{x})$ again differs from underlying ground truth for the target domain $P_{\mathcal{T}}(\mathbf{y}|\mathbf{x})$ and can be seen as an imperfect annotators. Our objective in this setting is also to jointly model $P_{\mathcal{T}}(\mathbf{y}, \mathbf{x})$ while noticing that there are significant domain shifts between $\mathcal{T}$ and any other $\mathcal{X}^{(k)}$.

4 Consensus Network

In this section, we present our two-phase framework ConNet for multi-source sequence labeling. As shown in Figure 2, our proposed framework first uses a multi-task learning schema with a special objective to decouple annotator representations as different parameters of a transformation around CRF layers. This **decoupling phase** (Section 4.2) is for decoupling the model parameters into a set of annotator-invariant model parameters and a set of annotator-specific representations. Secondly, the dynamic **aggregation phase** (Section 4.3) learns to contextually utilize the annotator representations with a lightweight attention mechanism to find the best suitable transformation for each sentence, so that the model can achieve a context-aware consensus among all sources. The inference process is described in Section 4.4.

4.1 The Base Model: BLSTM-CRF

Many recent sequence labeling frameworks (Ma and Hovy, 2016b; Misawa et al., 2017) share a very basic structure: a bidirectional LSTM network followed by a CRF tagging layer (i.e. BLSTM-CRF). The BLSTM encodes an input sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ into a sequence of hidden state vectors $\mathbf{h}_{1:n}$. The CRF takes as input the hidden state vectors and computes an emission score matrix $\mathbf{U} \in \mathbb{R}^{n \times L}$ where L is the size of tag set. It also maintains a trainable transition matrix $\mathbf{M} \in \mathbb{R}^{L \times L}$. We can consider $\mathbf{U}_{i,j}$ is the score of labeling the tag with id $j \in \{1, 2, ..., L\}$ for i^{th}



prediction ($\hat{\mathbf{y}}^{(k)}$)
CRF
BLSTM
sentence ($\mathbf{x}_i^k$)
Annotator $\{\mathbf{A}^{(k)}\}$
source ID (k)
Decoupling Phase
Weighted Voting
prediction ($\hat{\mathbf{y}}$)
Attention ($\mathbf{Q}$)
Consensus $\mathbf{A}_i^*$
CRF
BLSTM
sentence ($\mathbf{x}$)
Aggregation Phase

Figure 2: **Overview of the ConNet framework.** The decoupling phase constructs the shared model (yellow) and source-specific matrices (blue). The aggregation phase dynamically combines crowd components into a consensus representation (blue) by a context-aware attention module (red) for each sentence x.

word in the input sequence $\mathbf{x}$, and $\mathbf{M}_{i,j}$ means the transition score from i^{th} tag to j^{th}.

The CRF further computes the score s for a predicted tag sequence $\mathbf{y} = \{y_1, y_2, ..., y_k\}$ as

$$s(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} (\mathbf{U}_{t,y_t} + \mathbf{M}_{y_{t-1},y_t}), \qquad (1)$$

and then tag sequence $\mathbf{y}$ follows the conditional distribution

$$P(\mathbf{y}|\mathbf{x}) = \frac{\exp s(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{y} \in Y_{\mathbf{x}}} \exp s(\mathbf{x}, \mathbf{y})}. \qquad (2)$$

4.2 The Decoupling Phase: Learning annotator representations

For decoupling annotator-specific biases in annotations, we represent them as a transformation on emission scores and transition scores respectively. Specifically, we learn a matrix $\mathbf{A}^{(k)} \in \mathbb{R}^{L \times L}$ for each imperfect annotator k and apply this matrix as transformation on $\mathbf{U}$ and $\mathbf{M}$ as follows:

$$s^{(k)}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} \left((\mathbf{U}\mathbf{A}^{(k)})_{t,y_t} + (\mathbf{M}\mathbf{A}^{(k)})_{y_{t-1},y_t} \right). \qquad (3)$$

From this transformation, we can see that the original score function s in Eq. 1 becomes an source-specific computation. The original emission and transformation score matrix $\mathbf{U}$ and $\mathbf{M}$ are still shared by all the annotators, while they both are transformed by the matrix $\mathbf{A}^{(k)}$ for k-th annotator. While training the model parameters in this phase, we follow a multi-task learning schema. That is, we share the model parameters for BLSTM and CRF (including $\mathbf{W}$, $\mathbf{b}$, $\mathbf{M}$), while updating $\mathbf{A}^{(k)}$ only by examples in $\mathcal{S}_k = \{\mathcal{X}^{(k)}, \mathcal{Y}^{(k)}\}$.

The learning objective is to minimize the negative log-likelihood of all source annotations:

$$\mathcal{L} = -\log \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{X}^{(k)}|} P(\mathbf{y}_i^{(k)} | \mathbf{x}_i^{(k)}) \ , \qquad (4)$$

$$P(\mathbf{y}_i^{(k)} | \mathbf{x}_i^{(k)}) = \frac{\exp s^{(k)}(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)})}{\sum_{\mathbf{y}'} \exp s^{(k)}(\mathbf{x}, \mathbf{y}')}. \qquad (5)$$

The assumption on the annotation representation $\mathbf{A}^{(k)}$ is that it can model the pattern of annotation bias. Each annotator can be seen as a noisy version of the shared model. For the k-th annotator, $\mathbf{A}^{(k)}$ models noise from labeling the current word and transferring from the previous label. Specifically, each entry $\mathbf{A}_{i,j}^{(k)}$ captures the probability of mistakenly labeling i-th tag to j-th tag. In other words, the base sequence labeling model in Sec. 4.1 learns the basic consensus knowledge while annotator-specific components add their understanding to predictions.

4.3 The Aggregation Phase: Dynamically Reaching Consensus

In the second phase, our proposed network learns a context-aware attention module for a consensus representation supervised by combined predictions on the target data. For each sentence in target data $\mathcal{T}$, these predictions are combined by weighted voting. The weight of each source is its normalized F_1 score on the training set. Through weighted voting on such augmented labels over all source sentences $\mathcal{X}$, we can find a good approximation of underlying truth labels.

For better generalization and higher speed, an attention module is trained to estimate the relevance of each source to the target under the supervision of generated labels. Specifically, we compute the sentence embedding by concatenating the last hidden states of the forward LSTM and the backward LSTM, *i.e.* $\mathbf{h}^{(i)} = [\overrightarrow{\mathbf{h}}_T^{(i)}; \overleftarrow{\mathbf{h}}_0^{(i)}]$. The attention module inputs the sentence embedding and outputs a normalized weight for each source:

$$\mathbf{q}_i = \text{softmax}(\mathbf{Q}\mathbf{h}^{(i)}), \quad \text{where } \mathbf{Q} \in \mathbb{R}^{K \times 2d}. \qquad (6)$$

where d is the size of each hidden state $\mathbf{h}^{(i)}$. Source-specific matrices $\{\mathbf{A}^{(k)}\}_{k=1}^{K}$ are then aggregated into a consensus representation $\mathbf{A}_i^*$ for sentence $\mathbf{x}_i \in \mathcal{X}$ by

$$\mathbf{A}_i^* = \sum_{k=1}^{K} q_{i,k}\mathbf{A}^{(k)}. \tag{7}$$

In this way, the consensus representation contains more information about sources which are more related to the current sentence. It also alleviates the contradiction problem among sources, because it could consider multiple sources of different emphasis. Since only an attention model with weight matrix $\mathbf{Q}$ is required to be trained, the amount of computation is relatively small. We assume the base model and annotator representations are well-trained in the previous phase. The main objective in this phase is to learn how to select most suitable annotators for the current sentence.

4.4 Parameter Learning and Inference

CONNET learns parameters through two phases described above. In the decoupling phase, each instance from source S_k is used for training the base sequence labeling model and its representation $\mathbf{A}^{(k)}$. In the aggregation phase, we use aggregated predictions from the first phase to learn a lightweight attention module. For each instance in the target corpus $\mathbf{x}_i \in \mathcal{T}$, we calculate its embedding $\mathbf{h}_i$ from BLSTM hidden states. With these sentence embeddings, the context-aware attention module assigns weight $\mathbf{q}_i$ to each source and dynamically aggregates source-specific representations $\{\mathbf{A}^{(k)}\}$ for inferring $\hat{\mathbf{y}}_i$. In the inference process, only the consolidated consensus matrix $\mathbf{A}_i^*$ is applied to the base sequence learning model. In this way, more specialist knowledge helps to deal with more complex instances.

4.5 Model Application

The proposed model can be applied to two practical multi-sourcing settings: learning with crowd annotations and unsupervised cross-domain model adaptation. In the crowd annotation learning setting, the training data of the same domain is annotated by multiple noisy annotators, and each annotator is treated as a source. In the decoupling phase, the model is trained on noisy annotations, and in the aggregation phase, it is trained with combined predictions on the training set. In the cross-domain setting, the model has access to unlabeled training data of the target domain and clean labeled data of multiple source domains. Each domain is treated as a source. In the decoupling phase, the model is trained on source domains, and in the aggregation phase, the model is trained on combined predictions on the training data of the target domain. Our framework can also extend to new tasks other than sequence labeling and cope with different encoders. We will demonstrate this ability in experiments.

Our method is also incorporated as a feature for controlling the quality of crowd-annotation in annotation frameworks such as AlpacaTag (Lin et al., 2019) and LEAN-LIFE (Lee et al., 2020).

5 Experiments

We evaluate CONNET in the two aforementioned settings of multi-source learning: learning with crowd annotations and unsupervised cross-domain model adaptation. Additionally, to demonstrate the generalization of our framework, we also test our method on sequence labeling with transformer encoder in Appendix B and text classification with MLP encoder in Section 5.5.

5.1 Datasets

Crowd-Annotation Datasets. We use crowd-annotation datasets based on the 2003 CoNLL shared NER task (Tjong Kim Sang and De Meulder, 2003). The real-world datasets, denoted as AMT, are collected by Rodrigues et al. (2014) using Amazon's Mechanical Turk where F1 scores of annotators against the ground truth vary from 17.60% to 89.11%. Since there is no development set in AMT, we also follow Nguyen et al. (2017) to use the AMT training set and CoNLL 2003 development and test sets, denoted as AMTC. Overlapping sentences are removed in the training set, which is ignored in that work. Additionally, we construct two sets of simulated datasets to investigate the quality and quantity of annotators. To simulate the behavior of a non-expert annotator, a CRF model is trained on a small subset of training data and generates predictions on the whole set. Because of the limited size of training data, each model would have a bias to certain patterns.

Cross-Domain Datasets. In this setting, we investigate three NLP tasks: POS tagging, NER and text classification. For POS tagging task, we use the GUM portion (Zeldes, 2017) of Universal Dependencies (UD) v2.3 corpus with 17 tags and 7

Methods	AMTC			AMT		
	Precision(%)	Recall(%)	F1-score(%)	Precision(%)	Recall(%)	F1-score(%)
CONCAT-SLM	**85.95**(±1.00)	57.96(±0.26)	69.23(±0.13)	**91.12**(±0.57)	55.41(±2.66)	68.89(±1.92)
MVT-SLM	84.78(±0.66)	62.50(±1.36)	71.94(±0.66)	86.96(±1.22)	58.07(±0.11)	69.64(±0.31)
MVS-SLM	84.76(±0.50)	61.95(±0.32)	71.57(±0.04)	86.95(±1.12)	56.23(±0.01)	68.30(±0.33)
DS-SLM (Nguyen et al., 2017)	72.30*	61.17*	66.27*	-	-	-
HMM-SLM (Nguyen et al., 2017)	76.19*	66.24*	70.87*	-	-	-
MTL-MVT (Wang et al., 2018)	81.81(±2.34)	62.51(±0.28)	70.87(±1.06)	88.88(±0.25)	65.04(±0.80)	75.10(±0.44)
MTL-BEA (Rahimi et al., 2019)	85.72(±0.66)	58.28(±0.43)	69.39(±0.52)	77.56(±2.23)	67.23(±0.72)	72.01(±0.85)
CRF-MA (Rodrigues et al., 2014)	-	-	-	49.40*	**85.60***	62.60*
Crowd-Add (Nguyen et al., 2017)	85.81(±1.53)	62.15(±0.18)	72.09(±0.42)	89.74(±0.10)	64.50(±1.48)	75.03(±1.02)
Crowd-Cat (Nguyen et al., 2017)	85.02(±0.98)	62.73(±1.10)	72.19(±0.37)	89.72(±0.47)	63.55(±1.20)	74.39(±0.98)
CL-MW (Rodrigues and Pereira, 2018)	-	-	-	66.00*	59.30*	62.40*
ConNet (Ours)	84.11(±0.71)	**68.61**(±0.03)	**75.57**(±0.27)	88.77(±0.25)	72.79(±0.04)	**79.99**(±0.08)
Gold (Upper Bound)	89.48(±0.32)	89.55(±0.06)	89.51(±0.21)	92.12(±0.31)	91.73(±0.09)	91.92(±0.21)

Table 1: **Performance on real-world crowd-sourced NER datasets.** The best score in each column excepting Gold is marked **bold**. * indicates number reported by the paper.

domains: academic, bio, fiction, news, voyage, wiki, and interview. For NER task, we select the English portion of the OntoNotes v5 corpus (Hovy et al., 2006). The corpus is annotated with 9 named entities with data from 6 domains: broadcast conversation (bc), broadcast news (bn), magazine (mz), newswire (nw), pivot text (pt), telephone conversation (tc), and web (web). Multi-Domain Sentiment Dataset (MDS) v2.0 (Blitzer et al., 2007) is used for text classification, which is built on Amazon reviews from 4 domains: books, dvd, electronics, and kitchen. Since the dataset only contains word frequencies for each review without raw texts, we follow the setting in Chen and Cardie (2018) considering 5,000 most frequent words and use the raw counts as the feature vector for each review.

5.2 Experiment Setup

For sequence labeling tasks, we follow Liu et al. (2018) to build the BLSTM-CRF architecture as the base model. The dimension of character-level, word-level embeddings and LSTM hidden layer are set as 30, 100 and 150 respectively. For text classification, each review is represented as a 5000-d vector. We use an MLP with a hidden size of 100 for encoding features and a linear classification layer for predicting labels. The dropout with a probability of 0.5 is applied to the non-recurrent connections for regularization. The network parameters are updated by stochastic gradient descent (SGD). The learning rate is initialized as 0.015 and decayed by 5% for each epoch. The training process stops early if no improvements in 15 continuous epochs and selects the best model on the development set. For the dataset without a development set, we report the performance on the 50-th epoch. For each experiment, we report the average performance and standard variance of 3 runs with different random initialization.

5.3 Compared Methods

We compare our models with multiple baselines, which can be categorized in two groups: wrapper methods and joint models. To demonstrate the theoretical upper bound of performance, we also train the base model using ground-truth annotations in the target domain (Gold).

A wrapper method consists of a label aggregator and a deep learning model. These two components could be combined in two ways: (1) aggregating labels on crowd-sourced training set then feeding the generated labels to a Sequence Labeling Model (SLM) (Liu et al., 2017); (2) feeding multi-source data to a Multi-Task Learning (MTL) (Wang et al., 2018) model then aggregating multiple predicted labels. We investigate multiple label aggregation strategies. CONCAT considers all crowd annotations as gold labels. MVT does majority voting on the token level, *i.e.*, the majority of labels $\{\mathbf{y}^k_{i,j}\}$ is selected as the gold label for each token $\mathbf{x}_{i,j}$. MVS is conducted on the sequence level, addressing the problem of violating Begin/In/Out (BIO) rules. DS (Dawid and Skene, 1979), HMM (Nguyen et al., 2017) and BEA (Rahimi et al., 2019) induce consensus labels with probability models.

In contrast with wrapper methods, joint models incorporate multi-source data within the structure of sequential taggers and jointly model all individual annotators. CRF-MA models CRFs with Multiple Annotators by EM algorithm (Rodrigues et al., 2014). Nguyen et al. (2017) augments the LSTM

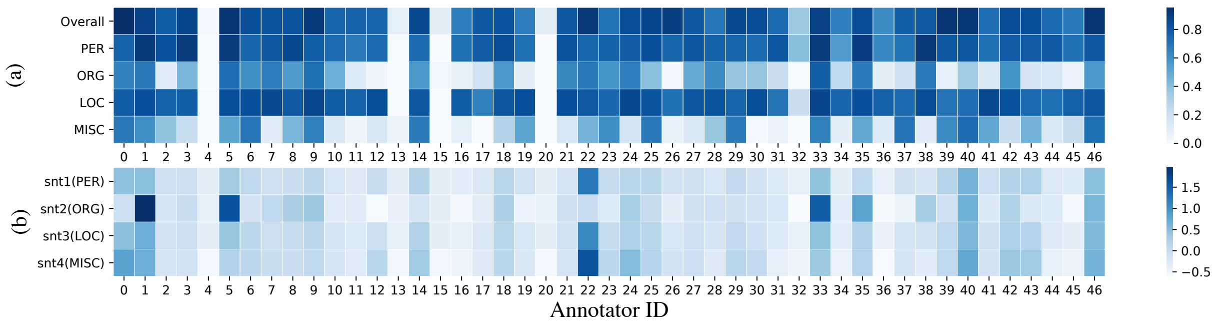


Figure 3: **Visualizations of (a) the expertise of annotators; (b) attention weights for sample sentences.** More cases and details are described in Appendix A.1.

architecture with crowd vectors. These crowd components are element-wise added to tags scores (`Crowd-Add`) or concatenated to the output of hidden layer (`Crowd-Cat`). These two methods are the most similar to our decoupling phase. We implemented them and got better results than reported. `CL-MW` applies a crowd layer to a CNN-based deep learning framework (Rodrigues and Pereira, 2018). `Tri-Training` uses bootstrapping with multi-task Tri-Training approach for unsupervised one-to-one domain adaptation (Saito et al., 2017; Ruder and Plank, 2018).

5.4 Learning with Crowd Annotations

Performance on real-world datasets. Tab. 1 shows the performance of aforementioned methods and our ConNet on two real-world datasets, *i.e.* AMT and AMTC[2]. We can see that ConNet outperforms all other methods on both datasets significantly on $F1$ score, which shows the effectiveness of dealing with noisy annotations for higher-quality labels. Although `CONCAT-SLM` achieves the highest precision, it suffers from low recall. Most existing methods have the high-precision but low-recall problem. One possible reason is that they try to find the latent ground truth and throw away illuminating annotator-specific information. So only simple mentions can be classified with great certainty while difficult mentions fail to be identified without sufficient knowledge. In comparison, ConNet pools information from all annotations and focus on matching knowledge to make predictions. It makes the model be able to identify more mentions and get a higher recall.

Case study. It is enlightening to analyze whether the model decides the importance of annotators given a sentence. Fig. 3 visualizes test F1 score of all annotators, and attention weights $\mathbf{q}_i$ in Eq. 6 for 4 sampled sentences containing different entity types. Obviously, the 2nd sample sentence with `ORG` has higher attention weights on 1st, 5th and 33rd annotator who are best at labeling `ORG`. More details and cases are shown in Appendix A.1.

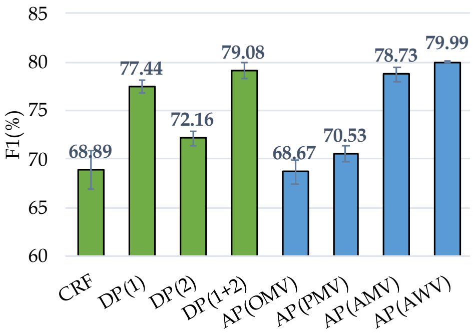
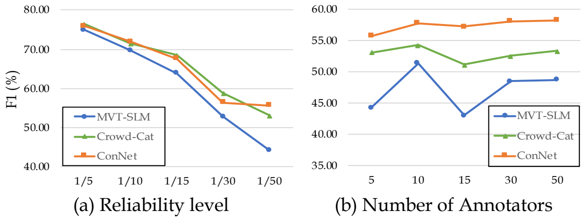


Figure 4: **Performance of ConNet variants** of decoupling phase (DP) and aggregation phase (AP).



(a) Reliability level

(b) Number of Annotators

Figure 5: **Performance on simulated crowd-sourced NER data** with (a) 5 annotators with different reliability levels; (b) different numbers of annotators with reliability $r = 1/50$.

Ablation study. We also investigate multiple variants of two phases on AMT dataset, shown in Fig. 4. We explore 3 approaches to incorporate source-specific representation in the decoupling phase (DP). `CRF` means the traditional approach as Eq. 1 while `DP(1+2)` is for our method as Eq. 3. `DP(1)` only applies source representations $\mathbf{A}^{(k)}$ to the emission score $\mathbf{U}$ while `DP(2)` only transfers the transition matrix $\mathbf{M}$. We can observe from the result that both variants can improve the result. The underlying model keeps more consensus knowledge if we extract annotator-specific bias on sentence encoding and label transition. We also compare 4 methods of generating supervision targets in the aggregation phase (AP). `OMV` uses ma-

[2] We tried our best to re-implement the baseline methods for all datasets, and left the results blank when the re-implementation is not showing consistent results as in the original papers.

Task & Corpus	**Multi-Domain POS Tagging: Universal Dependencies v2.3 - GUM**							
Target Domain	academic	bio	fiction	news	voyage	wiki	interview	AVG *Acc*(%)
CONCAT	92.68	92.12	93.05	90.79	92.38	92.32	91.44	92.11(±0.07)
MTL-MVT (Wang et al., 2018)	92.42	90.59	91.16	89.69	90.75	90.29	90.21	90.73(±0.29)
MTL-BEA (Rahimi et al., 2019)	92.87	91.88	91.90	91.03	91.67	91.31	91.29	91.71(±0.06)
Crowd-Add (Nguyen et al., 2017)	92.58	91.91	91.50	90.73	91.74	90.47	90.61	91.36(±0.14)
Crowd-Cat (Nguyen et al., 2017)	92.71	91.71	92.48	91.15	92.35	91.97	91.22	91.94(±0.08)
Tri-Training (Ruder and Plank, 2018)	92.84	92.15	92.51	91.40	92.35	91.29	91.00	91.93(±0.01)
CONNET	92.97	92.25	93.15	91.06	92.52	**92.74**	91.66	92.33(±0.17)
Gold (Upper Bound)	92.64	93.10	93.15	91.33	93.09	94.67	92.20	92.88(±0.14)

Task & Corpus	**Multi-Domain NER: OntoNotes v5.0 - English**						
Target Domain	nw	wb	bn	tc	bc	mz	AVG F_1(%)
CONCAT	68.23	32.96	77.25	53.66	72.74	62.61	61.24(±0.92)
MTL-MVT (Wang et al., 2018)	65.74	33.25	76.80	53.16	69.77	63.91	60.44(±0.45)
MTL-BEA (Rahimi et al., 2019)	58.33	32.62	72.47	47.83	48.99	52.68	52.15(±0.58)
Crowd-Add (Nguyen et al., 2017)	45.76	32.51	50.01	26.47	52.94	28.12	39.30(±4.44)
Crowd-Cat (Nguyen et al., 2017)	68.95	32.61	78.07	53.41	**74.22**	65.55	62.14(±0.89)
Tri-Training (Ruder and Plank, 2018)	69.68	33.41	79.62	47.91	70.85	68.53	61.67(±0.31)
CONNET	**71.31**	**34.06**	79.66	52.72	71.47	**70.71**	**63.32**(±0.81)
Gold (Upper Bound)	84.70	46.98	83.77	52.57	73.05	70.58	68.61(±0.64)

Task & Corpus	**Multi-Domain Text Classification: MDS**				
Target Domain	books	dvd	electronics	kitchen	AVG *Acc*(%)
CONCAT	75.68	77.02	81.87	83.07	79.41(±0.02)
MTL-MVT (Wang et al., 2018)	74.92	74.43	79.33	81.47	77.54(±0.06)
MTL-BEA (Rahimi et al., 2019)	74.88	74.60	79.73	82.82	78.01(±0.28)
Crowd-Add (Nguyen et al., 2017)	75.72	77.35	81.25	82.90	79.30(±9.21)
Crowd-Cat (Nguyen et al., 2017)	76.45	77.37	81.22	83.12	79.54(±0.25)
Tri-Training (Ruder and Plank, 2018)	77.58	78.45	81.95	83.17	80.29(±0.02)
CONNET	**78.75**	**81.06**	**84.12**	83.45	**81.85**(±0.04)
Gold (Upper Bound)	78.78	82.11	86.21	85.76	83.22(±0.19)

Table 2: **Performance on cross-domain data** The best score (except the Gold) in each column that is significantly ($p < 0.05$) better than the second best is marked **bold**, while those are better but not significantly are underlined.

jority voting of original annotations, while PMV substitutes them with model prediction learned from DP. AMV extends the model by using all prediction, while AWV uses majority voting weighted by each annotator's training $F1$ score. The results show the effectiveness of AWV, which could augment training data and well approximate the ground truth to supervise the attention module for estimating the expertise of annotator on the current sentence. We can also infer labels on the test set by conducting AWV on predictions of the underlying model with each annotator-specific components. However, it leads to heavy computation-consuming and unsatisfying performance, whose test $F1$ score is $77.35(\pm 0.08)$. We can also train a traditional BLSTM-CRF model with the same AMV labels. Its result is $78.93(\pm 0.13)$, which is lower than CONNET and show the importance of extracted source-specific components.

Performance on simulated datasets. To analyze the impact of annotator quality, we split the origin train set into z folds and each fold could be used to train a CRF model whose reliability could be represented as $r = 1/z$ assuming a model with less training data would have stronger bias and less generalization. We tried 5 settings where $z = \{5, 10, 15, 30, 50\}$ and randomly select 5 folds for each setting. When the reliability level is too low, *i.e.* $1/50$, only the base model is used for prediction without annotator representations. Shown in Fig. 5(a), CONNET achieves significant improvements over MVT-SLM and competitive performance as Crowd-Cat, especially when annotators are less reliable.

Regarding the annotator quantity, we split the train set into 50 subsets ($r = 1/50$) and randomly select $\{5, 10, 15, 30, 50\}$ folds for simulation. Fig. 5(b) shows CONNET is superior to baselines and able to well deal with many annotators while there is no obvious relationship between the performance and annotator quantity in baselines. We can see the performance of our model

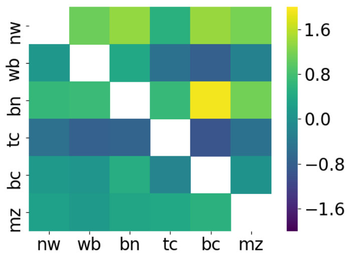


Figure 6: **Heatmap of averaged attention scores** from each source domain to each target domain.

increases as the number of annotators and, regardless of the number of annotators, our method consistently outperforms than other baselines.

5.5 Cross-Domain Adaptation Performance

The results of each task on each domain are shown in Tab. 2. We can see that CONNET performs the best on most of the domains and achieves the highest average score for all tasks. We report the accuracy for POS tagging and classification, and the chunk-level $F1$ score for NER. We can see that CONNET achieves the highest average score on all tasks. `MTL-MVT` is similar to our decoupling phase and performs much worse. Naively doing unweighted voting does not work well.

The attention can be viewed as implicitly doing weighted voting on the feature level. `MTL-BEA` relies on a probabilistic model to conduct weighted voting over predictions, but unlike our approach, its voting process is independent from the input context. It is probably why our model achieves higher scores. This demonstrates the importance of assigning weights to domains based on the input sentence.

`Tri-Training` trains on the concatenated data from all sources also performs worse than CONNET, which suggests the importance of a multi-task structure to model the difference among domains. The performance of `Crowd-Add` is unstable (high standard deviation) and very low on the NER task, because the size of the crowd vectors is not controllable and thus may be too large. On the other hand, the size of the crowd vectors in `Crowd-Cat` can be controlled and tuned. These two methods leverage domain-invariant knowledge only but not domain-specific knowledge and thus does not have comparable performance.

5.6 Analyzing Learned Attention

We analyzed the attention scores generated by the attention module on the OntoNotes dataset. For each sentence in the target domain we collected the attention score of each source domain, and finally the attention scores are averaged for each source-target pair. Fig. 6 shows all the source-to-target average attention scores. We can see that some domains can contribute to other related domains. For example, bn (broadcast news) and nw (newswire) are both about news and they contribute to each other; bn and bc (broadcast conversation) are both broadcast and bn contributes to bc; bn and nw both contributes to mz (magzine) probably because they are all about news; wb (web) and tc (telephone conversation) almost make no positive contribution to any other, which is reasonable because they are informal texts compared to others and they are not necessarily related to the other. Overall the attention scores can make some sense. It suggests that the attention is aware of relations between different domains and can contribute to the model.

6 Conclusion

In this paper, we present CONNET for learning a sequence tagger from multi-source supervision. It could be applied in two practical scenarios: learning with crowd annotations and cross-domain adaptation. In contrast to prior works, CONNET learns fine-grained representations of each source which are further dynamically aggregated for every unseen sentence in the target data. Experiments show that our model is superior to previous crowd-sourcing and unsupervised domain adaptation sequence labeling models. The proposed learning framework also shows promising results on other NLP tasks like text classification.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007, United States Office Of Naval Research under Contract No. N660011924033, and NSF SMA 18-29268. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. We would like to thank all the collaborators in USC INK research lab for their constructive feedback on the work.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of ACL*, pages 440–447.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proc. of NAACL-HLT*, pages 1226–1240, New Orleans, Louisiana. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proc. of EMNLP*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Dredze, Partha Pratim Talukdar, and Koby Crammer. 2009. Sequence learning from data with multiple labels. In *ECML/PKDD Workshop on Learning from Multi-Label Data*.

Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. 2015. Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recognition Letters*, 65:60–66.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. 2017. Who said what: Modeling individual labelers improves classification. In *AAAI*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90\% solution. In *Proc. of NAACL-HLT*.

John Lafferty. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289. Morgan Kaufmann.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of NAACL-HLT*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Ouyu Lan, Su Zhu, and Kai Yu. 2018. Semi-supervised training using adversarial multi-task learning for spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6049–6053. IEEE.

Dong-Ho Lee, Rahul Khanna, Bill Yuchen Lin, Jamin Chen, Seyeon Lee, Qinyuan Ye, Elizabeth Boschee, Leonardo Neves, and Xiang Ren. 2020. Leanlife: A label-efficient annotation framework towards learning from explanation. In *Proc. of ACL (Demo)*.

Bill Y Lin, Frank Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proc. of ACL Workshop*, pages 160–165.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. Triggerner: Learning with entity triggers as explanations for named entity recognition. In *ACL*.

Bill Yuchen Lin, Dongho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren. 2019. Alpacatag: An active learning-based crowd annotation framework for sequence tagging. In *Proc. of ACL (Demo)*.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proc. of EMNLP*.

Liyuan Liu, Xiang Ren, Jingbo Shang, Jian Peng, and Jiawei Han. 2018. Efficient Contextualized Representation: Language Model Pruning for Sequence Labeling. In *Proc. of EMNLP*.

Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, and Jiawei Han. 2017. Heterogeneous supervision for relation extraction: A representation learning approach. In *Proc. of EMNLP*, pages 46–56. Association for Computational Linguistics.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. 2014. Transfer joint matching for unsupervised domain adaptation. In *Proc. of CVPR*, pages 1410–1417.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to chinese word segmentation. In *Proc. of SIGHAN Workshop*.

Xuezhe Ma and Eduard Hovy. 2016a. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proc. of ACL*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016b. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proc. of ACL*, pages 1064–1074. Association for Computational Linguistics.

Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. 2015. Unsupervised domain adaptation with imbalanced cross-domain data. In *Proc. of ICCV*, pages 4121–4129.

Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2017. Character-based bidirectional LSTM-CRF with words and characters for Japanese named entity recognition. In *Proc. of ACL Workshop*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proc. of ACL*, pages 299–309. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. In *Proc. of ACL*, pages 151–164.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. of EMNLP*.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher R. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. Sequence labeling with multiple annotators. *Machine Learning*, 95(2):165–181.

Filipe Rodrigues and Francisco C Pereira. 2018. Deep learning from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proc. of ACL*, pages 1044–1054, Melbourne, Australia. Association for Computational Linguistics.

Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *Proc. of ICML*, pages 2988–2997. JMLR. org.

Edwin Simpson, Jonas Pfeiffer, and Iryna Gurevych. 2020. Low resource sequence tagging with weak labels. In *AAAI 2020*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. of NAACL-HLT*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, page bty869.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. Crossweigh: Training named entity tagger from imperfect annotations. In *EMNLP/IJCNLP*.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proc. of NAACL-HLT*, pages 672–682.

Amir Zeldes. 2017. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Analysis of ConNet with BLSTM Encoder

A.1 Case study on learning with crowd annotations

To better understand the effect and benefit of CONNET, we do some case study on AMTC real-world dataset with 47 annotators. We look into some more instances to investigate the ability of attention module to find right annotators in Fig. 7 and Tab. 3. Sentence 1-12 contains a specific entity type respectively while 13-16 contains multiple different entities. Compared with expertise of annotators, we can see that the attention module would give more weight on annotators who have competitive performance and preference on the included entity type. Although top selected annotators for ORG has relatively lower expertise on ORG than PER and LOC, they are actually the top three annotators with highest expertise on ORG.

B Result of ConNet with Transformer Encoder

To demonstrate the generalization of our framework, we re-implement CONNET and some baselines (MTV-SLM, Crowd-Add, Gold) with Transformer-CRF as the base model. Specifically, the base model takes Transformer as the encoder for CRF, which has shown its effectiveness in many NLP tasks (Vaswani et al., 2017; Devlin et al., 2019). Transformer models sequences with self-attention and eliminates all recurrence. Following the experimental settings from (Vaswani et al., 2017), we set the number of heads for multi-head attention as 8, the dimension of keys and values as 64, and the hidden size of the feed-forward layers as 1024. We conduct experiments with crowd-annotation dataset AMTC on NER task and cross-domain dataset UD on POS task, which are described in Section 5.1. Results are shown in Table 4. We can see our model outperforms over other baselines in both tasks and applications.

1	Defender [PER Hassan Abbas] rose to intercept a long ball into the area in the 84th minute but only managed to divert it into the top corner of [PER Bitar] ’s goal .
2	[ORG Plymouth] 4 [ORG Exeter] 1
3	Hosts [LOC UAE] play [LOC Kuwait] and [LOC South Korea] take on [LOC Indonesia] on Saturday in Group A matches .
4	The former [MISC Soviet] republic was playing in an [MISC Asian Cup] finals tie for the first time .
5	[PER Bitar] pulled off fine saves whenever they did .
6	[PER Coste] said he had approached the player two months ago about a comeback .
7	[ORG Goias] 1 [ORG Gremio] 3
8	[ORG Portuguesa] 1 [ORG Atletico Mineiro] 0
9	[LOC Melbourne] 1996-12-06
10	On Friday for their friendly against [LOC Scotland] at [LOC Murrayfield] more than a year after the 30-year-old wing announced he was retiring following differences over selection .
11	Scoreboard in the [MISC World Series]
12	Cricket - [MISC Sheffield Shield] score .
13	“ He ended the [MISC World Cup] on the wrong note , ” [PER Coste] said .
14	Soccer - [ORG Leeds] ’ [PER Bowyer] fined for part in fast-food fracas .
15	[ORG Rugby Union] - [PER Cuttitta] back for [LOC Italy] after a year .

Table 3: Sample instances in Fig. 3 and Fig. 7 with NER annotations including PER (red), ORG (blue), LOC (violet) and MISC (orange).

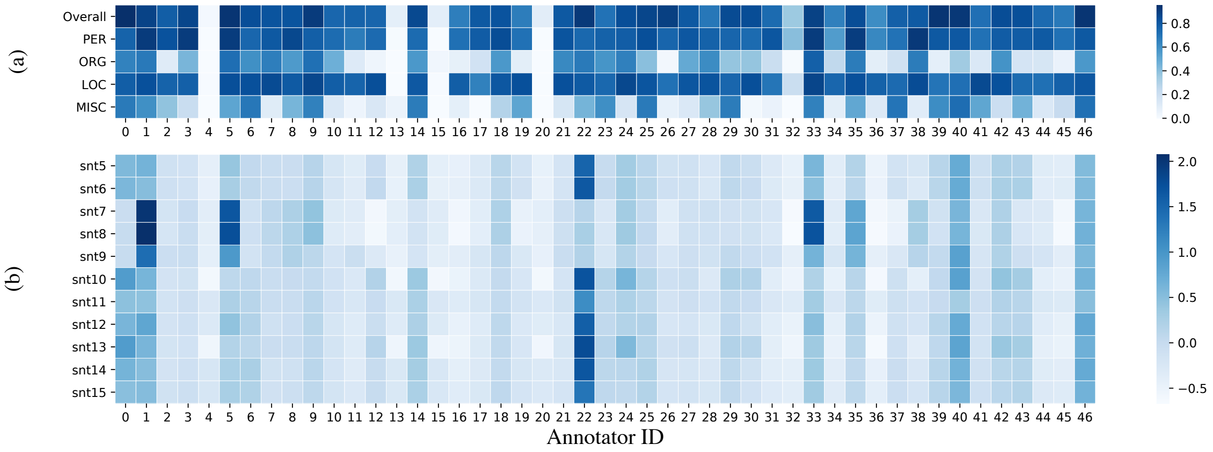


Figure 7: Visualizations of (a) the expertise of annotators; (b) attention weights for additional sample sentences to Fig. 3. Details of samples are described in Tab. 3.

Methods	**AMTC**			**UD**
	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)
MVT-SLM	72.21(±1.63)	51.72(±3.58)	60.21(±1.87)	87.23(±0.51)
Crowd-Add (Nguyen et al., 2017)	75.32(±1.41)	50.80(±0.30)	60.68(±0.67)	88.20(±0.36)
ConNet (Ours)	**76.86**(±0.33)	**56.43**(±3.32)	**65.05**(±2.32)	**89.27**(±0.31)
Gold (Upper Bound)	81.24(±1.25)	80.52(±0.37)	80.87(±0.79)	90.45(±0.71)

Table 4: Performance of methods with Transformer-CRF as the base model on crowd-annotation NER dataset AMTC and cross-domain POS dataset UD.