PABI: A UNIFIED PAC-BAYESIAN INFORMATIVENESS MEASURE FOR INCIDENTAL SUPERVISION SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world applications often require making use of *a range of incidental supervision signals*. However, we currently lack a principled way to measure the benefit an incidental training dataset can bring, and the common practice of using indirect, weaker signals is through exhaustive experiments with various models and hyper-parameters. This paper studies whether we can, *in a single framework, quantify the benefit of various types of incidental signals for one's target task without going through combinatorial experiments*. We propose PABI, a unified informativeness measure backed by PAC-Bayesian theory, characterizing the reduction in uncertainty that indirect, weak signals provide. We demonstrate PABI's use in quantifying various types of incidental signals including partial labels, noisy labels, constraints, cross-domain signals, and combinations of these. Experiments with various setups on two natural language processing (NLP) tasks, named entity recognition (NER) and question answering (QA), show that PABI correlates well with learning performance, providing a promising way to determine, ahead of learning, which supervision signals would be beneficial.

1 INTRODUCTION

The supervised learning paradigm, where direct supervision signals are assumed to be available in high-quality and large amounts, has been struggling to fulfill the needs in many real-world AI applications. As a result, researchers and practitioners often resort to datasets that are not collected directly for the target task but, hopefully, capture some phenomena useful for it (Pan & Yang, 2009; Vapnik & Vashist, 2009; Roth, 2017; Kolesnikov et al., 2019). However, it remains unclear how to predict the benefits of these incidental signals on our target task beforehand, so the common practice is often trial-and-error: do experiments with different combinations of datasets and learning protocols, often exhaustively, to achieve improvement on a target task (Liu et al., 2019; Khashabi et al., 2020). Not only this is very costly, this trial-and-error approach can also be hard to interpret: *if we don't see improvements, is it because the incidental signals themselves are not useful for our target task, or is it because the learning protocols we have tried are inappropriate?*

The difficulties of foreshadowing the benefits of various incidental supervision signals are two-fold. First, it is hard to provide a *unified* measure because of the intrinsic differences among different signals (e.g., how do we predict and compare the benefit of learning from noisy data and the benefit of knowing some constraints for the target task?). Second, it is hard to provide a *practical* measure supported by *theory*. Previous attempts are either not practical or too heuristic (Baxter, 1998; Ben-David et al., 2010; Thrun & O'Sullivan, 1998; Gururangan et al., 2020). In this paper, we propose a unified PAC-Bayesian based informativeness measure (PABI) to quantify the value of incidental signals. We suggest that the informativeness of various incidental signals can be uniformly characterized by the reduction in the original concept class uncertainty they provide. Specifically, in the PAC-Bayesian framework, the informativeness is based on the Kullback–Leibler (KL) divergence between the prior and the posterior, where incidental signals are used to estimate a better prior (closer to the gold posterior) to achieve better generalization performance. Furthermore, we provide a more practical entropy-based approximation of PABI.

We have been in need of a unified informativeness measure like PABI. For instance, it might be obvious that we can expect better learning performance if the training data are less noisy and more completely annotated, but what if we want to compare the benefits of a noisy dataset and that of a

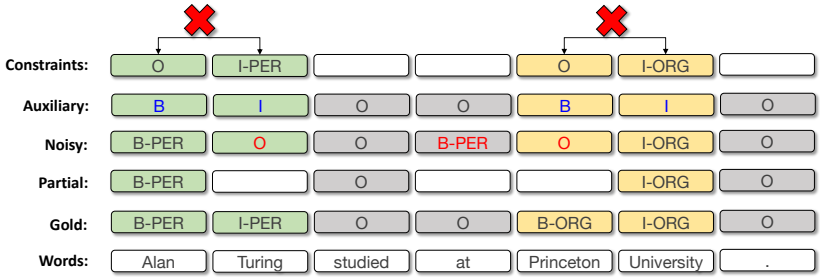


Figure 1: An example of NER with various incidental supervision signals: partial labels (some missing labels in structured outputs), noisy labels (some incorrect labels), auxiliary labels (labels of another task, e.g. named entity detection in the figure), and constraints in structured learning (e.g. the BIO constraint where I-X must follow B-X or I-X (Ramshaw & Marcus, 1999) in the figure).

partial dataset? PABI enables this kind of comparisons beforehand, on a wide range of incidental signals such as partial labels, noisy labels, constraints, auxiliary signals, cross-domain signals, and some combinations of them, for sequence tagging tasks in NLP. A specific example of NER is shown in Fig. 1, and the advantages of PABI are in Table 1.

Finally, our experiments on two NLP tasks, NER and QA, show that there is a strong positive correlation between PABI and the relative improvement for various incidental signals. This strong positive correlation indicates that the proposed unified, theory-motivated measure PABI can serve as a good indicator of the final learning performance, providing a promising way to know which signals are helpful for a target task beforehand.

Organization. We start with related work in Section 1.1. Then we derive informativeness measure PABI in Section 2. We show examples on how to compute PABI using various incidental signals in Section 3. We verify the effectiveness of PABI in Section 4. Section 5 concludes this paper.

1.1 Related Work

There are lots of practical measures proposed to quantify the benefits of specific types of signals. For example, a widely used measure for partial signals in structured learning is the partial rate (Cour et al., 2011; Hovy & Hovy, 2012; Liu & Dietterich, 2014; Van Rooyen & Williamson, 2017; Ning et al., 2019); a widely used measure for for noisy signals is the noise ratio (Angluin & Laird, 1988; Natarajan et al., 2013; Rolnick et al., 2017; Van Rooyen & Williamson, 2017); Ning et al. (2019) propose to use the concaveness of the mutual information with different percentage of annotations to quantify the strength of constraints in the structured learning; others, in NLP, have quantified the contribution of constraints experimentally (Chang et al., 2012; 2008). Bjerva (2017) proposes to use conditional entropy or mutual information to quantify the value for auxiliary signals. As for domain adaptation, domain similarity can be measured by the performance gap between domains (Wang et al., 2019) or measures based on the language model in NLP, such as the vocabulary overlap (Gururangan et al., 2020). Among them, the most relevant work is (Bjerva, 2017). However, their conditional entropy or mutual information is based on token-level label distribution, which cannot be used for incidental signals involving multiple tokens or inputs, such as constraints and cross-domain signals. The key advantage of our proposed measure PABI is that PABI is a unified measure supported by the PAC-Bayesian theory for a broader range of incidental signals compared to these practical measures for specific types of incidental signals.

There also has been a line of theoretical work that attempts to exploit incidental supervision signals. Among them, the most relevant part is task relatedness. Ben-David & Borbely (2008) define the task relatedness based on the richness of transformations between inputs for different tasks, but their analysis is limited to cases where data is from the same classification problem but the inputs are in different subspace. Juba (2006) proposes to use the joint Kolmogorov complexity (Li et al., 2008) to characterize relatedness, but it is still unclear how to compute the joint Kolmogorov complexity in real-world applications. Mahmud & Ray (2008) further propose to use conditional Kolmogorov complexity to measure the task relatedness and provide empirical analysis for decision trees, but it is unclear how to use their relatedness for other models, such as deep neural networks. Thrun & O'Sullivan (1998) propose to cluster tasks based on the similarity between the task-optimal distance

	Incidental Supervision Signals					**Unified Measure**		Support	
Measures	partial	noisy	constraints	auxiliary	cross-domain	**cross-type**	**mixed-type**	Theoretical	Empirical
CST'11, HH'12, LD'14	✓	×	×	×	×	×	×	✓	✓
AL'88, NDRT'13, RVBS'17	×	✓	×	×	×	×	×	✓	✓
VW'17, WNR'20	✓	✓	×	×	×	×	×	✓	×
NHFR'19	✓	×	✓	×	×	×	×	×	✓
B'17	×	×	×	✓	×	×	×	×	✓
GMSLBDS'20	×	×	×	×	✓	×	×	×	✓
PABI (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between PABI and prior works: (1) Cross-type: PABI is a unified measure which can measure the benefit of different types of incidental signals (e.g., comparing a noisy dataset and a partially annotated dataset). (2) Mixed-type: PABI can measure the benefit of mixed incidental signals (e.g., a dataset that is both noisy and partially annotated). (3) PABI is derived from PAC-Bayesian theory but also easy to compute in practice; PABI is shown to have similar or better predicting capability of signals' benefit (see Figs. 2 and 3 and Sec. 3.1).[2]

metric of k-nearest neighbors (KNN), but their analysis is based on KNN and it is unclear how to use their relatedness for other models. A lot of other works provide quite good qualitative analysis to show various incidental signals are helpful but they did not provide quantitative analysis to quantify to what extent these types of incidental signals can help (Balcan & Blum, 2010; Abu-Mostafa, 1993; Natarajan et al., 2013; Van Rooyen & Williamson, 2017; Baxter, 1998; London et al., 2016; Ciliberto et al., 2019; Ben-David et al., 2010; Wang et al., 2020). Compared to these theoretical analyses, PABI can be easily used in practice to quantify the benefits of a broader range of incidental signals.

2 PABI: A Unified PAC-Bayesian Informativeness Measure

We start with notations and preliminaries. Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the label space, and $\hat{\mathcal{Y}}$ be the prediction space. Let $\mathcal{D}$ denote the underlying distribution on $\mathcal{X} \times \mathcal{Y}$. Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ be the loss function that we use to evaluate learning algorithms. A set of training samples $S = \{x_i, y_i\}_{i=1}^m$ is generated i.i.d. from $\mathcal{D}$. In the common supervised learning setting, we usually assume the concept that generates data comes from the concept class $\mathcal{C}$. In this paper, we assume $\mathcal{C}$ is finite, and its size is $|\mathcal{C}|$, which is common in NLP. We want to choose a predictor $c : \mathcal{X} \to \hat{\mathcal{Y}}$ from $\mathcal{C}$ such that it generalizes well to unseen data with respect to ℓ, measured by the *generalization error* $R_{\mathcal{D}}(c) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}}[\ell(\mathbf{y}, c(\mathbf{x}))]$. The *training error* over S is $R_S(c) = \frac{1}{m}\sum_{i=1}^m \ell(y_i, c(x_i))$.

More generally, instead of predicting a concept, we can specify a distribution over the concept class. Let $\mathcal{P}$ denote the space of probability distributions over $\mathcal{C}$. General Bayesian learning algorithms (Zhang et al., 2006) in the PAC-Bayesian framework (McAllester, 1999a;b; Seeger, 2002; McAllester, 2003b;a; Maurer, 2004; Guedj, 2019) aim to choose a posterior $\pi_\lambda \in \mathcal{P}$ over the concept class $\mathcal{C}$ based on a prior $\pi_0 \in \mathcal{P}$ and training data S, where λ is a hyper parameter that controls the tradeoff between the prior and the data likelihood. In this setting, the training error and the generalization error need to be generalized, to take the distribution into account, as $L_S(\pi_\lambda) = \mathbb{E}_{c\sim\pi_\lambda}[R_S(c)]$ and $L_{\mathcal{D}}(\pi_\lambda) = \mathbb{E}_{c\sim\pi_\lambda}[R_{\mathcal{D}}(c)]$ respectively. One can easily see that when the posterior is *one-hot* (exactly one entry of the distribution is 1), we have the original definitions of training error and generalization error, as in the PAC framework (Valiant, 1984).

2.1 Informativeness Measures in the PAC-Bayesian Framework

We are now ready to derive the proposed informativeness measure PABI based on PAC-Bayes. The generalization error bound in the PAC-Bayesian framework (Guedj, 2019; Catoni, 2007) says that with probability $1-\delta$ over S, $L_{\mathcal{D}}(\pi_{\lambda^*}) \leq L_{\mathcal{D}}(\pi^*) + \sqrt{\frac{8B(D_{KL}(\pi^*||\pi_0) + \ln\frac{2\ln(mC)}{\delta})}{m}}$, where π_{λ^*} is the posterior distribution with the optimal $\lambda^* = \sqrt{\frac{2m(D_{KL}(\pi^*||\pi_0)+\ln\frac{2}{\delta})}{B}}$, $\pi^* \in \mathcal{P}$ is the gold posterior that generates the data, $D_{KL}(\pi^*||\pi_0)$ denotes the KL divergence from π_0 to π^*, B and C are two constants. This is based on the Theorem 2 in Guedj (2019).

[2]Papers are denoted by abbreviations of author names as: CST'11 (Cour et al., 2011), HH'12 (Hovy & Hovy, 2012), LD'14 (Liu & Dietterich, 2014), AL'88 (Angluin & Laird, 1988), NDRT'13 (Natarajan et al., 2013), RVBS'17 (Rolnick et al., 2017), VW'17 (Van Rooyen & Williamson, 2017), WNR'20 (Wang et al., 2020), NHFR'19 (Ning et al., 2019), B'17 (Bjerva, 2017), GMSLBDS'20 (Gururangan et al., 2020).

As shown in the generalization bound, the generalization error is bounded by the KL divergence $D_{KL}(\pi^*||\pi_0)$ from the prior distribution to the gold posterior distribution. *Therefore, we propose to utilize incidental signals to improve the prior distribution from π_0 to $\tilde{\pi}_0$ so that it is closer to the gold posterior distribution π^*.* Correspondingly, we can define PABI, the informativeness measure for incidental supervision signals, by measuring the improvement with regard to the gold posterior, in KL divergence sense.

Definition 2.1 (PABI)**.** *Suppose we use incidental signals to improve the prior distribution from π_0 to $\tilde{\pi}_0$. The informativeness measure for incidental signals, PABI, is defined as*

$$S(\pi_0, \tilde{\pi}_0) \triangleq \sqrt{1 - \frac{D_{KL}(\pi^*||\tilde{\pi}_0)}{D_{KL}(\pi^*||\pi_0)}} \tag{1}$$

Remark. Note that $S(\pi_0, \tilde{\pi}_0) = 0$ if $\tilde{\pi}_0 = \pi_0$, while if $\tilde{\pi}_0 = \pi^*$, then $S(\pi_0, \tilde{\pi}_0) = 1$. This result is consistent with our intuition that the closer $\tilde{\pi}_0$ is to π^*, the more benefits we can gain from incidental signals. The square root function is used in PABI for two reasons: first, the generalization bounds in both PAC-Bayesian and PAC (see Sec. 2.2) frameworks have the square root function; second, in our later experiments, we find that square root function can significantly improve the Pearson correlation between the relative performance improvement and PABI.

In our paper, we focus on the setting that the gold posterior π^* is *one-hot*, which means π^* concentrates on the true concept $c^* \in \mathcal{C}$, though the definition of PABI can handle general gold posterior. However, π^* is unknown in practice, which makes Eq. (1) hard to be computed in reality. In the following, we provide an approximation $\hat{S}$ of PABI.

Definition 2.2 (Approximation of PABI)**.** *Assume that the original prior π_0 is uniform, and the gold posterior π^* is one-hot concentrated on the true concept c^* in $\mathcal{C}$, as we have assumed that $\mathcal{C}$ is finite. Let $H(\cdot)$ be the entropy function. The approximation $\hat{S}$ of PABI is defined as*

$$\hat{S}(\pi_0, \tilde{\pi}_0) \triangleq \sqrt{1 - \frac{H(\tilde{\pi}_0)}{H(\pi_0)}} = \sqrt{1 - \frac{H(\tilde{\pi}_0)}{\ln|\mathcal{C}|}} \tag{2}$$

The uniform prior π_0 is usually used when we do not have information about the prior on which concept in the class that generates data. *The intuition behind $\hat{S}$ is that, it measures how much entropy incidental signals reduce, compared with non-informative prior π_0.* $\hat{S}$ can be computed through data and thus is practical. To see how this approximation works, first note $D_{KL}(\pi^*||\pi_0) = \ln|\mathcal{C}|$ because π^* is one-hot and π_0 is uniform over the finite concept class $\mathcal{C}$. Let π_c be the one-hot distribution concentrated on concept c for each $c \in \mathcal{C}$. The approximation is that we estimate π^* by π_c, where c follows $\tilde{\pi}_0$: $D_{KL}(\pi^*||\tilde{\pi}_0) \approx \mathbb{E}_{c\sim\tilde{\pi}_0} D_{KL}(\pi_c||\tilde{\pi}_0)$. It turns out $\mathbb{E}_{c\sim\tilde{\pi}_0} D_{KL}(\pi_c||\tilde{\pi}_0) = H(\tilde{\pi}_0)$. Therefore,

$$\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{H(\tilde{\pi}_0)}{\ln|\mathcal{C}|}} = \sqrt{1 - \frac{\mathbb{E}_{c\sim\tilde{\pi}_0} D_{KL}(\pi_c||\tilde{\pi}_0)}{\ln|\mathcal{C}|}} \approx \sqrt{1 - \frac{D_{KL}(\pi^*||\tilde{\pi}_0)}{D_{KL}(\pi^*||\pi_0)}} = S(\pi_0, \tilde{\pi}_0)$$

2.2 PABI IN THE PAC FRAMEWORK

We have derived PABI in the PAC-Bayesian framework. Here, we discuss briefly on what PABI reduces to in the PAC framework and what limitations are when PABI is restricted to the PAC framework. The generalization bound in the PAC framework (Mohri et al., 2018) says with probability $1-\delta$ over S, $R_{\mathcal{D}}(c) \leq R_S(c) + \sqrt{\frac{\ln|\mathcal{C}| + \ln\frac{2}{\delta}}{2m}}$. *We propose to reduce the concept class from $\mathcal{C}$ to $\tilde{\mathcal{C}}$ by using incidental signals.* Then PABI in the PAC framework can be written as

$$S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}} \tag{3}$$

It turns out that Eq. (3) is a special case of Eq. (1) when π^* is one-hot over $\mathcal{C}$, π_0 is uniform over $\mathcal{C}$ and $\tilde{\pi}_0$ is uniform over $\tilde{\mathcal{C}}$ (See Appx. A.1 for the derivation).

However, PABI restricted to the PAC framework cannot handle the probabilistic cases. For example, incidental signals can reduce the probability of some concepts, though the concept class is not reduced. In this example, $S(\mathcal{C}, \tilde{\mathcal{C}})$ is zero, but we actually benefit from incidental signals. Some

analysis on the extensions and general limitations of PABI can be found in Appx. A.2 and A.3. We need to notice that the size of concept class also plays an important role in the lower bound on the generalization error (more details in Appx. A.4), indicating that PABI based on the reduction of the concept class is a reasonable measure.

3 EXAMPLES FOR PABI

In this section, we show some examples of sequence tagging tasks[3] in NLP for PABI. Similar to the categorization of transfer learning (Pan & Yang, 2009), we use *inductive signals* to denote the signals with a different conditional probability distribution ($P(\mathbf{y} \mid \mathbf{x})$) from gold signals, such as noisy and auxiliary signals, and *transductive signals* to denote the signals with the same task ($P(\mathbf{y} \mid \mathbf{x})$) as gold signals but a different marginal distribution of $\mathbf{x}$ ($P(\mathbf{x})$) from gold signals, such as cross-domain and cross-lingual signals. In our following analysis, we study the tasks with finite concept class which is quite common in NLP. For simplicity, we focus on simple cases where the number of incidental signals is large enough. How different factors (including base model performance, size of incidental signals, data distributions, algorithms, cost-sensitive losses) affect PABI are discussed in Appx. A.5. We derive the PABI for partial labels in detail and the derivations for others are similar. More examples and details can be found in Appx. A.6.

3.1 EXAMPLES WITH INDUCTIVE SIGNALS

Partial labels. The labels for each example in sequence tagging tasks is a sequence and some of them are unknown in this case. Assuming that the ratio of the unknown labels in data is $\eta_p \in [0, 1]$, the size of the reduced concept class will be $|\tilde{\mathcal{C}}| = |\mathcal{L}|^{n|\mathcal{V}|^n \eta_p}$. Therefore, $\hat{S}(\pi_0, \tilde{\pi}_0) = S(\pi_0, \tilde{\pi}_0) = S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}} = \sqrt{1 - \frac{|\mathcal{L}|^{n|\mathcal{V}|^n \eta_p}}{|\mathcal{L}|^{n|\mathcal{V}|^n}}} = \sqrt{1 - \eta_p}$. It is consistent with the widely used partial rate because it is a monotonically decreasing function of the partial rate.

Noisy labels. For each token, $P(y|\tilde{y})$ is determined by the noisy rate $\eta_n \in [0, 1]$, i.e. $P(y = \tilde{y}) = 1 - \eta_n$ and the probability of other labels are all $\frac{\eta_n}{|\mathcal{L}|-1}$. We can get the corresponding probability distribution of labels over the tokens in all inputs ($\tilde{\pi}_0$ over the concept class). In this way, $\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{\eta_n \ln(|\mathcal{L}|-1) - \eta_n \ln \eta_n - (1-\eta_n)\ln(1-\eta_n)}{\ln|\mathcal{L}|}}$. It is consistent with the widely used noise rate because it is a monotonically decreasing function of the noisy rate. In practice, the noisy rate can be easily estimated with some aligned data[4], and the noise with more complex patterns (e.g. input dependent) is postponed as our future work.

3.2 EXAMPLES WITH TRANSDUCTIVE SIGNALS

For transductive signals, such as cross-domain signals, we can first extend the concept class $\mathcal{C}$ to the extended concept class $\mathcal{C}^e$ with the corresponding extended input space $\mathcal{X}^e$. After that, we can use incidental signals to estimate a better prior distribution $\tilde{\pi}_0^e$ over the extended concept class $\mathcal{C}^e$, and then get the corresponding $\tilde{\pi}_0$ over the original concept class by restricting the concept from $\mathcal{X}^e$ to $\mathcal{X}$. In this way, the informativeness of transductive signals can still be measured by $S(\pi_0, \tilde{\pi}_0)$ or $\hat{S}(\pi_0, \tilde{\pi}_0)$. The restriction step is similar to Roth & Zelenko (2000).

However, how to compute $H(\tilde{\pi}_0)$ is still unclear. We now provide a way to estimate it. For simplicity, we use $c(\mathbf{x})$ to denote the gold system on the gold signals, $\tilde{c}(\mathbf{x})$ to denote the perfect system on the incidental signals, and $\hat{c}(\mathbf{x})$ to denote the silver system trained on the incidental signals. Source domain (target domain) is the domain of incidental signals (gold signals). We use $\mathcal{D}$ to denote the target domain and $\tilde{\mathcal{D}}$ to denote the source domain. $P_{\mathcal{D}}(\mathbf{x})$ is the marginal distribution of $\mathbf{x}$ under $\mathcal{D}$, and similar definition for $P_{\tilde{\mathcal{D}}}(\mathbf{x})$. In our analysis, we assume $\tilde{c}(\mathbf{x})$ is a noisy version of $c(\mathbf{x})$ with

[3] Given an input $\mathbf{x} \in \mathcal{X} = \mathcal{V}^n$ generated from a distribution $\mathcal{D}$, the task aims to get the corresponding label $\mathbf{y} \in \hat{\mathcal{Y}} = \mathcal{Y} = \mathcal{L}^n$, where $\mathcal{V}$ is the vocabulary of input words, $\mathcal{L}$ is the label set for the task, and n is the length of the input sentence.

[4] If there is no jointly annotated data, we can use similar methods as Bjerva (2017) to create some approximate jointly annotated data.

the noisy rate η, and $\hat{c}(\mathbf{x})$ is a noisy version of $\tilde{c}(\mathbf{x})$ with the noisy rate η_1 (η_1') in the source (target) domain : $\eta_1 = \mathbb{E}_{\mathbf{x}\sim P_{\tilde{\mathcal{D}}}(\mathbf{x})}\mathbf{1}(\hat{c}(\mathbf{x}) \neq \tilde{c}(\mathbf{x}))$ ($\eta_1' = \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}(\mathbf{x})}\mathbf{1}(\hat{c}(\mathbf{x}) \neq \tilde{c}(\mathbf{x}))$).

In practice, η is unknown but it can be estimated by η_1 in the source domain and $\eta_2 = \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}(\mathbf{x})}\mathbf{1}(\hat{c}(\mathbf{x}) \neq c(\mathbf{x}))$ (the noisy rate of the silver system compared to the gold system on the target domain) as follows:

$$\eta = \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}(\mathbf{x})}\mathbf{1}(c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})) = \frac{(|\mathcal{L}|-1)(\eta_1'-\eta_2)}{1-|\mathcal{L}|(1-\eta_1')} = \frac{(|\mathcal{L}|-1)(\eta_1-\eta_2)}{1-|\mathcal{L}|(1-\eta_1)}. \tag{4}$$

Here we add an assumption: η_1' in the target domain is equal to η_1 in the source domain. The derivation of Eq. (4) can be found in Appx A.7. Similar to noisy labels, the corresponding informativeness of transductive signals can be then computed as $\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{\eta \ln(|\mathcal{L}|-1) - \eta \ln \eta - (1-\eta)\ln(1-\eta)}{\ln|\mathcal{L}|}}$.

To justify the use of η in the informativeness measure for transductive signals, we show in Theorem A.2 (see Appx. A.8) that (informally speaking) the generalization error of a learner that is jointly trained on data from both source and target domains can be upper bounded by η (plus a function of the size of the concept class and the number of samples). Finally we note that although the computation cost of PABI for transductive signals is higher than that for inductive signals, it is still much cheaper than building combined models with joint training.

3.3 Examples with Mixed Incidental Signals

The mix of partial and noisy labels. The corresponding informativeness for the mix of partial and noisy labels is $\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{(1-\eta_p) * (1 - \frac{\eta_n \ln(|\mathcal{L}|-1) - \eta_n \ln \eta_n - (1-\eta_n)\ln(1-\eta_n)}{\ln|\mathcal{L}|})}$, where $\eta_p \in [0, 1]$ denotes the ratio of unlabeled tokens, and $\eta_n \in [0, 1]$ denotes the noise ratio.

The mix of partial labels and constraints. For BIO constraints with partial labels, we can use dynamic programming with sampling as Ning et al. (2019) to estimate $\ln|\tilde{\mathcal{C}}|$ and $\hat{S}(\pi_0, \tilde{\pi}_0)$.

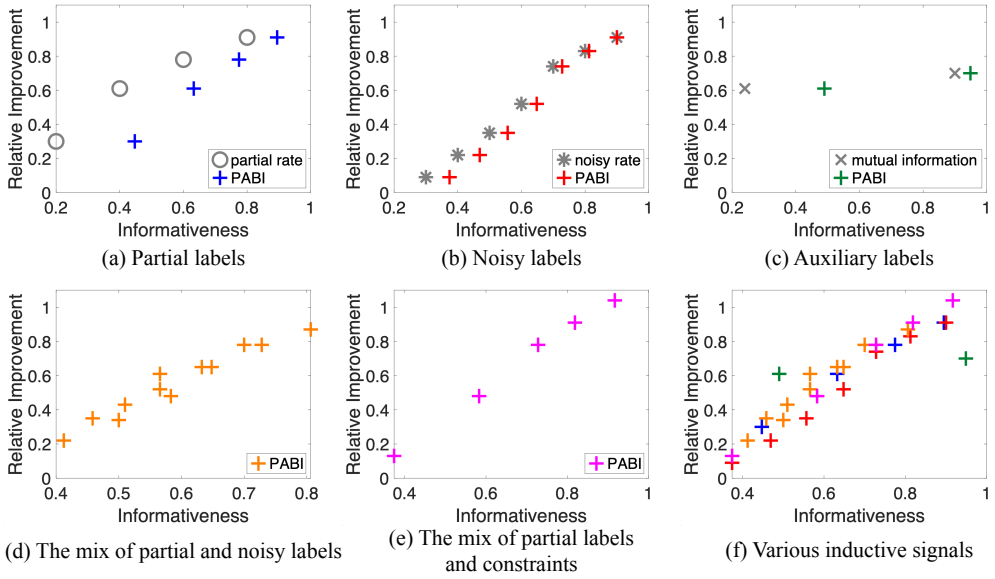
4 Experiments

In this section, we verify the effectiveness of PABI on various inductive signals and transductive signals on NER and QA. More details about experimental settings are in Appx. A.9.

Learning with various inductive signals. In this part, we analyze the informativeness of inductive signals for NER. We use Ontonotes NER (18 types of named entities) (Hovy et al., 2006) as the main task. We randomly sample 10% sentences (30716 words) of the development set as the small gold signals, 90% sentences (273985 words) of the development set as the large incidental signals. We use a two-layer NNs with 5-gram features as our basic model. The lower bound for our experiments is the result of the model with small gold Ontonotes NER annotations and bootstrapped on the unlabeled texts of the large gold Ontonotes NER, which is 38 F1, and the upper bound is the result of the model with both small gold Ontonotes NER annotations and the large gold Ontonotes NER annotations, which is 61 F1. To utilize inductive signals, we propose a new bootstrapping based algorithm CWBPP (Algorithm 1 in Appx. A.10), where inductive signals are used to improve the inference stage by approximating a better prior. It is an extension of CoDL (Chang et al., 2007) with various inductive signals.

We experiment on NER with various inductive signals, including three types of individual signals, partial labels, noisy labels, auxiliary labels, and two types of mixed signals: signals with both partial and noisy labels, and signals with both partial labels and constraints. As shown in Fig. 2, we find that there is a strong correlation between the relative improvement and PABI for various inductive signals. For individual signals in Fig. 2(a)-2(c), we find that PABI have similar foreshadowing ability comparing to the measures for specific signals, i.e., $1-\eta_p$ for partial labels, $1-\eta_n$ for noisy labels, and entropy normalized mutual information[5] ($\frac{I(Y;\tilde{Y})}{H(Y)}$) for auxiliary labels. For mixed signals in Fig. 2(d)-2(e), the strong correlation is quite promising because the benefits of mixed signals

[5] In Bjerva (2017), they propose to use mutual information or conditional entropy to measure the informativeness, so we normalize the mutual information with the entropy to make the value between 0 and 1.



(a) Partial labels (b) Noisy labels (c) Auxiliary labels

(d) The mix of partial and noisy labels (e) The mix of partial labels and constraints (f) Various inductive signals

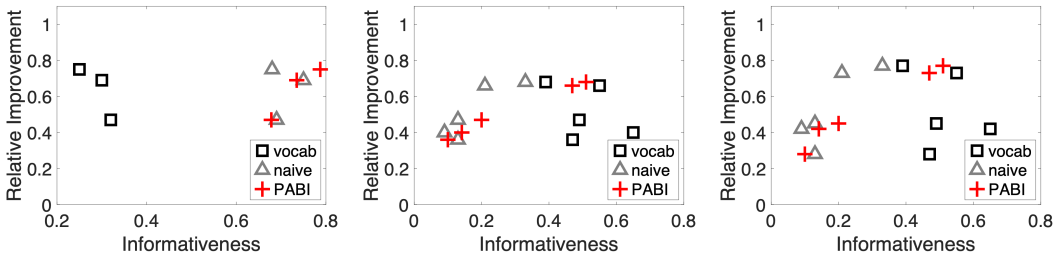
Figure 2: **Correlations between informativeness and relative performance improvement for NER with various inductive signals:** partial labels, noisy labels, auxiliary labels, mix of partial and noisy labels, and mix of partial labels and constraints. For NER with various inductive signals (f) (with all PABI points from (a)-(e)), Pearson's correlation and Spearman's rank correlation are 0.92 and 0.93. Note that the relative improvement for named entity detection (with informativeness 0.90 but relative improvement 0.70) in auxiliary labels (c) is smaller than expected mainly due to the **imbalanced label distribution** (88% O among all BIO labels). More discussions about the imbalanced distribution can be found in Appx. A.5. For individual inductive signals, we also show the results of baselines (gray points), i.e. one minus partial rate for partial labels, one minus noisy rate for noisy labels, and entropy normalized mutual information for auxiliary labels. We can see that PABI has similar foreshadowing ability with these measures for specific signals, and can measure the benefits of mixed inductive signals and compare different types of inductive signals.

cannot be quantified by existing frameworks. Finally, the strong positive correlation for different types of signals in Fig. 2(f) indicates that it is feasible to compare the benefits of different incidental signals with PABI, which cannot be addressed by existing frameworks.

Learning with cross-domain signals. In this part, we consider the benefits of cross-domain signals for NER and QA. For NER, we consider four NER datasets, Ontonotes, CoNLL, twitter(Strauss et al., 2016), and GMB (Bos et al., 2017). We aim to detect the person names here because the only shared type of the four datasets is the person[6]. In our experiments, the twitter NER serves as the main dataset and other three datasets are cross-domain datasets. There are 85 sentences in the small gold training set, 756 sentences (9 times of the gold signals) in the large incidental training set, and 851 sentences in the test set. For QA, we consider SQuAD (Rajpurkar et al., 2016), QAMR (Michael et al., 2017), Large QA-SRL (FitzGerald et al., 2018), QA-RE (Levy et al., 2017), NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017). In our experiments, the SQuAD dataset servers as the main dataset and other datasets are cross-domain datasets. We randomly sample 700 QA pairs as the small gold signals, about $6.2K$ QA pairs as the large incidental signals (9 times of the small gold signals), and $21K$ QA pairs as the test data.

We use BERT as our basic model and consider two strategies to make use of incidental signals: joint training and pre-training. For NER, the lower bound for is the result with only small gold twitter annotations, which is 61.51 F1, and the upper bound is the result with both small gold twitter annotations and large gold twitter annotations, which is 78.31. For QA, the lower bound is the result with only small gold SQuAD annotations, which is 26.45 exact match. The upper bound for the

[6]Note that our focus here is cross-domain signals, the divergent set of classes for different domains is a mix of cross-domain and auxiliary signals, which is our future work.



(a) Joint-training NER with baselines and PABI

(b) Joint-training QA with baselines and PABI

(b) Pre-training QA with baselines and PABI

Figure 3: **Correlation between informativeness measures (baselines or the PABI) and relative performance improvement** (via joint training or pre-training) for cross-domain NER and cross-domain QA. Red results with the PABI which is based on η in Eq. (4); Gray points indicate the results with the naive informativeness measure η_2; Black points indicate the results with the vocabulary overlap baseline as in (Gururangan et al., 2020). The Pearson's correlation of three informativeness measure in the three cases are: $0.96/0.19/-0.85$, $1.00/0.88/-0.40$, $0.99/0.85/-0.30$, indicating the quality of the PABI measure. Similarly, the corresponding Spearman's rank correlation are: $1.00/-0.50/-1.00$, $1.00/0.82/-0.30$, $1.00/0.82/-0.30$.

joint training is the result with both small gold SQuAD annotations and large SQuAD annotations, which is 50.72 exact match. Similarily, the upper bound for the pre-training is 49.24 exact match.

The relation between the relative improvement (pre-training or joint training) and informativeness measures (baselines or the PABI) is shown in Fig. 3. We can see that there is a strong positive correlation between the relative improvement and PABI for cross-domain signals. Comparing to the naive baseline η_2, we can see that the adjustment from η_1 is crucial (Eq. (4)), indicating that directly using η_2 is not a good choice. We also show the vocabulary overlap baseline as in (Gururangan et al., 2020) where we compute the overlap over the top $1K$ most frequent unigrams (excluding stop words and punctuations) between different domains. The results for this baseline are quite bad, and the fact that our data is not so large makes this baseline more valueless.

5 Conclusion and Future Work

Supported by PAC-Bayesian theory, this paper proposes a unified framework, PABI, to characterize incidental supervision signals by how much uncertainty they can reduce in the hypothesis space. We demonstrate the effectiveness of PABI in foreshadowing the benefits of various signals, i.e., partial labels, noisy labels, auxiliary labels, constraints, cross-domain signals and combinations of them, for solving NER and QA. To our best knowledge, PABI is the first informativeness measure that can handle various incidental signals and combinations of them; PABI is motivated from PAC-Bayes and can be easily computed in real-world tasks. As the recent success of natural language modeling has given rise to many explorations in knowledge transfer across tasks and corpora (Bjerva, 2017; Phang et al., 2018; Zhu et al., 2019; Liu et al., 2019; He et al., 2020; Khashabi et al., 2020) , PABI is a concrete step towards explaining some of these observations.

PABI can also provide guidance in designing learning protocols. For instance, in a B/I/O sequence chunking task,[7] missing labels make it a partial annotation problem, while treating missing labels as O introduces noise. Since the informativeness of partial signals is larger than that of noisy signals with the same partial/noisy rate (see details in Sec. 3.1), PABI suggests us *not* to treat missing labels as O, and this is exactly what Mayhew et al. (2019) prove to us via their experiments. We plan to explore more in this direction to apply PABI in designing better learning protocols.

Finally, we need to acknowledge that our current exploration for auxiliary labels is still limited. The results for auxiliary labels with a different label set (Fig. 2(c)) is blocked by the imbalanced label distribution (Appx. A.5). For more complex cases, such as part-of-speech tagging (PoS) for NER, we can only treat them as cross-sentence constraints now and the results are also limited (Appx. A.6). In future, we will work more in this direction to better quantify the value of auxiliary signals.

[7] B/I/O indicates if a token is the begin/inside/outside of a text span.

REFERENCES

Yaser S Abu-Mostafa. Hints and the VC dimension. *Neural Computation*, 5(2):278–288, 1993.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83, 2017.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Jonathan Baxter. Theoretical models of learning to learn. In *Learning to learn*, pp. 71–94. Springer, 1998.

Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Johannes Bjerva. Will my auxiliary tagging task help? estimating auxiliary tasks effectivity in multi-task learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pp. 216–220, 2017.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. The groningen meaning bank. In *Handbook of linguistic annotation*, pp. 463–496. Springer, 2017.

Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding Semi-Supervision with Constraint-Driven Learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 280–287, Prague, Czech Republic, 6 2007. Association for Computational Linguistics.

Ming-Wei Chang, Lev Ratinov, Nicholas Rizzolo, and Dan Roth. Learning and Inference with Constraints. 7 2008.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. Structured Learning with Constrained Conditional Models. *Machine Learning*, pp. 399–431, 6 2012.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems*, pp. 7354–7364, 2018.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in neural information processing systems*, pp. 4412–4420, 2016.

Carlo Ciliberto, Francis Bach, and Alessandro Rudi. Localized structured prediction. In *Advances in Neural Information Processing Systems*, pp. 7299–7309, 2019.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. Large-scale QA-SRL parsing. In *ACL*, pp. 2051–2060, 2018.

Benjamin Guedj. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, July 2020.

Hangfeng He, Qiang Ning, and Dan Roth. QuASE: Question-Answer Driven Sentence Encoding. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Dirk Hovy and Eduard Hovy. Exploiting partial annotations with em training. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pp. 31–38, 2012.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60, 2006.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, July 2017.

Brendan Juba. Estimating relatedness via data compression. In *Proceedings of the 23rd international conference on Machine learning*, pp. 441–448, 2006.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UnifiedQA: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *CoNLL*, pp. 333–342, 2017.

Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, pp. 1629–1637, 2014.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL-HLT*, pp. 1073–1094, 2019.

Ben London, Bert Huang, and Lise Getoor. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

MM Mahmud and Sylvian Ray. Transfer learning using kolmogorov complexity: Basic theory and empirical evaluations. In *Advances in neural information processing systems*, pp. 985–992, 2008.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. Named Entity Recognition with Partially Annotated Training Data. In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*, 2019.

David McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pp. 203–215. 2003a.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999a.

David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999b.

David A McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003b.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. Crowdsourcing question-answer meaning representations. *NAACL*, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pp. 1196–1204, 2013.

Qiang Ning, Hangfeng He, Chuchu Fan, and Dan Roth. Partial or Complete, That's The Question. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pp. 157–176. Springer, 1999.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

Dan Roth. Incidental Supervision: Moving beyond Supervised Learning. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, 2 2017.

Dan Roth and Dmitry Zelenko. Towards a theory of Coherent Concepts. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, pp. 639–644, 2000.

Erik Tjong Kim Sang and Fien De Meulder. 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*, pp. 142–147, 2003.

Matthias Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 138–144, 2016.

Sebastian Thrun and Joseph O'Sullivan. Clustering learning tasks and the selective cross-task transfer of knowledge. In *Learning to learn*, pp. 235–257. Springer, 1998.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200. Association for Computational Linguistics, August 2017.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Brendan Van Rooyen and Robert C Williamson. A theory of learning with corrupted labels. *The Journal of Machine Learning Research*, 18(1):8501–8550, 2017.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

Boyu Wang, Jorge Mendez, Mingbo Cai, and Eric Eaton. Transfer learning via minimizing the performance gap between domains. In *Advances in Neural Information Processing Systems*, pp. 10645–10655, 2019.

Kaifu Wang, Qiang Ning, and Dan Roth. Learnability with indirect supervision signals. *arXiv preprint arXiv:2006.08791*, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Tong Zhang et al. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

Xiaofeng Zhu, Feng Liu, Goce Trajcevski, and Dingding Wang. Frosting weights for better continual training. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 506–510. IEEE, 2019.