A APPENDIX

A.1 `PABI` IN THE PAC FRAMEWORK

Suppose π^* is one-hot over $\mathcal{C}$, π_0 is uniform over $\mathcal{C}$ and $\tilde{\pi}_0$ is uniform over $\tilde{\mathcal{C}}$. We have $D_{KL}(\pi^*||\pi_0) = \ln|\mathcal{C}|$ and $D_{KL}(\pi^*||\tilde{\pi}_0) = \ln|\tilde{\mathcal{C}}|$. It follows that

$$S(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{D_{KL}(\pi^*||\tilde{\pi}_0)}{D_{KL}(\pi^*||\pi_0)}} = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}} = S(\mathcal{C}, \tilde{\mathcal{C}}).$$

A.2 INFORMATIVENESS MEASURES IN PARAMETRIC CONCEPT CLASS

In practice, algorithms are often based on parametric concept class. The two informativeness measures in the PAC-Bayesian framework, $S(\pi_0, \tilde{\pi}_0)$ and $\hat{S}(\pi_0, \tilde{\pi}_0)$, can be easily adapted to handle the cases in parametric concept class. Given parametric space $\mathcal{C}_w$, we can easily change the probability distribution $\pi(\mathcal{C}_w)$ over the parametric concept class to the probability distribution $\pi(\mathcal{C})$ over the finite concept class $\mathcal{C} = \{c : \mathcal{V}^n \rightarrow \mathcal{L}^n\}$ by clustering concepts in the parametric space according to their outputs on all inputs. The concepts in each cluster have the same outputs on all inputs as outputs of one concept in the finite concept class $\mathcal{C}$. We then merge the probabilities of concepts in the same cluster to get the probability distribution $\pi(\mathcal{C})$ over the finite concept class $\mathcal{C}$. This merging approach can be applied to any concept class which is not equal to the finite concept class $\mathcal{C}$, including non-parametric and semi-parametric concept class. In practice, we can use sampling algorithms, such as Markov chain Monte Carlo (MCMC) methods, to simulate this clustering strategy.

A.3 LIMITATIONS OF INFORMATIVENESS MEASURES

Different informativeness measures are based on different assumptions, so we analyze their limitations in detail to understand their limitations in applications.

For the informativeness measure $S(\mathcal{C}, \tilde{\mathcal{C}})$, it cannot handle probabilistic signals or infinite concept classes. There are various probabilistic incidental signals, such as soft constraints and probabilistic co-occurrences between an auxiliary task and the main task. An example of probabilistic co-occurrences between part-of-speech (PoS) tagging and NER is that the adjectives have a 95% probability to have the label O in NER. As for the infinite concept class, most classifiers are based on infinite parametric spaces. Thus, $S(\mathcal{C}, \tilde{\mathcal{C}})$ cannot be applied to these classifiers.

The informativeness measure $S(\pi_0, \tilde{\pi}_0)$ is hard to be computed for some complex cases. In practice, we can use the estimated posterior distribution over the gold data, which is asymptotically unbiased, to estimate it. Another approximation is to use the informativeness measure $\hat{S} = \sqrt{1 - \frac{H(\tilde{\pi}_0)}{H(\pi_0)}}$. However, it is not directly linked to the generalization bound, so more work is needed to guarantee its reliability for some complex probabilistic cases. We postpone to provide the theoretical guarantees for $\hat{S} = \sqrt{1 - \frac{H(\tilde{\pi}_0)}{H(\pi_0)}}$ on more complex cases as our future work.

A.4 LOWER BOUND IN THE PAC FRAMEWORK

In the following theorem, we show that the VC dimension (size of concept class) also plays an important role in the lower bund for the generalization error, indicating that `PABI` based on the reduction of the concept class is a reasonable measure.

Theorem A.1. *Let $\mathcal{C}$ be a concept class with VC dimension $d > 1$. Then, for any $m \geq 1$ and any learning algorithm $\mathcal{A}$, there exists a distribution $\mathcal{D}$ over $\mathcal{X}$ and a target concept $c \in \mathcal{C}$ such that*

$$P_{S \sim \mathcal{D}^m}[R_{\mathcal{D}}(c_S) > \frac{d-1}{32m}] \geq 1/100$$

where c_S is a consistent concept with S returned by $\mathcal{A}$. This is the Theorem 3.20 in Chapter 3.4 of Mohri et al. (2018).

A.5 DISCUSSION OF SOME OTHER FACTORS IN `PABI`

In this subsection, we consider the impact of the following factors in `PABI`: base model performance, the size of incidental signals, data distribution, algorithm and cost-sensitive loss.

Base model performance. In the generalization bound in both PAC and PAC-Bayesian, we can see that the relative improvement in the generalization bound from reducing $\mathcal{C}$ is small if m is large. In practice, the relative improvement is the real improvement with some noise. Therefore, we can see that the real improvement is dominant if m is small and the noise is dominant if m is large. Therefore, `PABI` may not work well when m is large and when the performance on the target task is already good enough.

The size of incidental signals. Our previous analysis is based on a strong assumption that incidental signals are large enough (ideally $\tilde{m} \to \infty$) A more realistic `PABI` is based on $\tilde{\mathcal{C}}$ with $\tilde{m}$ examples as $S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}_{\tilde{m}}|}{\ln|\mathcal{C}|}} = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}_{\tilde{m}}|}{\ln|\mathcal{C}_{\tilde{m}}|} \times \frac{\ln|\mathcal{C}_{\tilde{m}}|}{\ln|\mathcal{C}|}}$, where $\mathcal{C}_{\tilde{m}}$ denotes the restricted concept class of $\mathcal{C}$ on the $\tilde{m}$ examples, and so does $\tilde{\mathcal{C}}_{\tilde{m}}$. (1) When $\tilde{m}$ is large enough, $S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}}$. (2) When the sizes of different incidental signals are all $\tilde{m}$, the relative improvement is independent of $\tilde{m}$ ($\frac{\ln|\tilde{\mathcal{C}}_{\tilde{m}}|}{\ln|\mathcal{C}_{\tilde{m}}|} = \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}$), and $S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|} \times \frac{\ln|\mathcal{C}_{\tilde{m}}|}{\ln|\mathcal{C}|}}$. Our experiments are based on this case and does not really rely on the assumption that incidental signals are large enough. (3) The incidental signals we are comparing are not large enough and have different sizes, we need to use $S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \tilde{m} * \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}}$ to incorporate that difference.

Data distribution. As for the distribution of examples, both PAC and PAC-Bayesian are distribution-free (see more in Chapter 2.1 of Mohri et al. (2018)). However, if we consider the joint distribution between examples and labels, such as imbalanced label distribution, the situation will be different. Specific types of joint data distribution refer to a restricted concept class $\mathcal{C}'$. Therefore, `PABI` is expected to work well if the reduction from $\mathcal{C}$ is similar to the reduction from $\mathcal{C}'$ with incidental signals, i.e. $S(\mathcal{C}', \tilde{\mathcal{C}}') = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}'|}{\ln|\mathcal{C}'|}} \approx \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}}$.

Algorithm. Different algorithms make different assumptions on the concept class. For example, SVM aims to find the maximum-margin hyperplane (see more in Chapter 5.4 of Mohri et al. (2018)). Therefore, a specific algorithm actually is based on a restricted concept class $\mathcal{C}'$ (e.g. concepts with margin in SVM case). Similarly, `PABI` is expected to work well if the reduction from $\mathcal{C}$ is similar to the reduction from $\mathcal{C}'$ with incidental signals. We also cannot compare the benefits from various incidental signals with different algorithms. If the algorithm is not expressive enough to take advantage of incidental signals, we may also not be able to use `PABI` there.

Cost-sensitive Loss. For different loss functions other than 0-1 loss, there are still some similar generalization bounds in PAC and PAC-Bayesian (using complexity of concept class and sample size) (Bartlett et al., 2006; Ciliberto et al., 2016). Therefore, `PABI` can also be used (possibly with some minor modifications) for cost-sensitive loss functions.

A.6 MORE EXAMPLES WITH INCIDENTAL SIGNALS

k-gram	1	2	3	4	5	6	7	8	9	10
word-pos	8.68	49.45	84.08	96.22	98.96	99.54	99.69	99.73	99.75	99.76
word-ner	27.65	76.23	92.98	98.04	99.37	99.74	99.84	99.88	99.89	99.90
pos-ner	0.20	6.65	13.78	25.36	41.50	60.14	77.04	88.61	95.01	97.92
ner-pos	0.00	0.01	0.03	0.07	0.17	0.39	0.80	1.47	2.45	3.71

Table 2: K-gram co-occurrence analysis for PoS and NER in the whole Ontonotes dataset. For example, word-pos represents the percentage of k-gram words that have the unique k-gram PoS labels.

In this subsection, we show more examples with incidental signals, including within-sentence constraints, cross-sentence constraints, auxiliary labels, cross-lingual signals, cross-modal signals, and the mix of cross-domian signals and constraints.

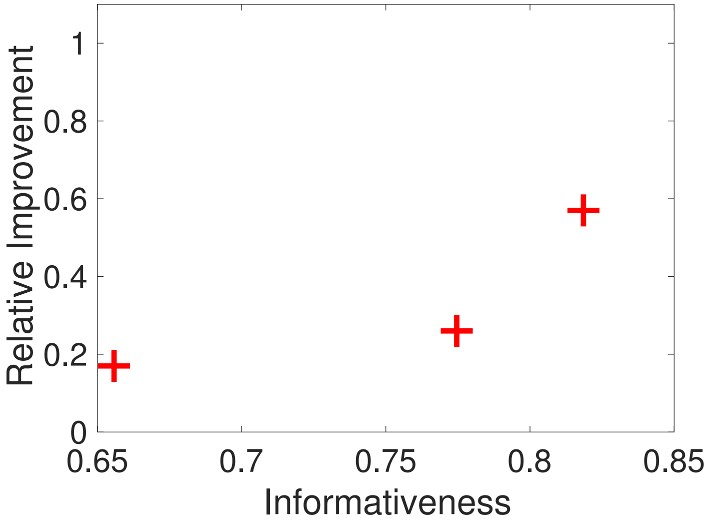

Figure 4: The correlations between the informativeness and the relative performance improvement for NER with cross-sentence constraints.

Within-Sentence Constraints. As for within-sentence constraints, we show three types of common constraints in NLP, which are BIO constraints, assignment constraints, and ranking constraints.

- BIO constraints are widely used in sequence tagging tasks, such as NER. For BIO constraints, I-X must follow B-X or I-X, where "X" is finer types such as PER (person) and LOC (location). We consider a simple case here: there are only B, I, O three labels. We have $\ln|\tilde{\mathcal{C}}| = |\mathcal{V}|^n(\ln|\mathcal{L}|^n + \ln[\sum_{m=0}^{\lfloor(n+1)/2\rfloor}\binom{m}{n-m+1}(\frac{-1}{|\mathcal{L}|^2})^m])$ for the BIO constraint. Therefore, $\hat{S}(\pi_0, \tilde{\pi}_0) = S(\pi_0, \tilde{\pi}_0) = S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\mathcal{L}|^n + \ln[\sum_{m=0}^{\lfloor(n+1)/2\rfloor}\binom{m}{n-m+1}(\frac{-1}{|\mathcal{L}|^2})^m]}{\ln|\mathcal{L}|^n}}$. This value can be approximated by the dynamic programming as Ning et al. (2019).
- Assignment constraints can be used in various types of semantic parsing tasks, such as semantic role labeling (SRL). Assume we need to assign d agents with d' tasks such that the agent nodes and the task nodes form a bipartite graph (without loss of generality, assume $d \leq d'$). Each agent is represented by a feature vector in $\mathcal{V}_f$. We have $\hat{S}(\pi_0, \tilde{\pi}_0) = S(\pi_0, \tilde{\pi}_0) = S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}} = \sqrt{1 - \frac{\ln\binom{d}{d'}}{d\ln d'}}$. This informativeness doesn't rely on the choice of $\mathcal{V}_f$ where that $\mathcal{V}_f$ denotes discrete feature space for arguments.
- Ranking constraints can be used in ranking problems, such as temporal relation extraction. For a ranking problem with t items, there are $d = t(t-1)/2$ pairwise comparisons in total. Its structure is a chain following the transitivity constraints, i.e., if $A < B$ and $B < C$, then $A < C$. In this way, we have $\hat{S}(\pi_0, \tilde{\pi}_0) = S(\pi_0, \tilde{\pi}_0) = S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}} = \sqrt{1 - \frac{\ln t!}{\ln 2^d}} \approx \sqrt{1 - \frac{2\ln t - 2}{(t-1)\ln 2}}$. This informativeness doesn't rely on the choice of $\mathcal{V}_f$ where $\mathcal{V}_f$ denotes discrete feature space for events.

Cross-sentence Constraints. For cross-sentence constraints, we consider a common example, global statistics based on 2-tuple of tokens, i.e. pairs of tokens in different sentences must have the same labels. We can group words into K groups with probability p. In this way, we have $\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{-p\ln p - (1-p)\ln(1-p) + p\ln|\mathcal{L}|^K + (1-p)\ln(|\mathcal{L}|^n|\mathcal{V}|^n - |\mathcal{L}|^K))}{\ln|\mathcal{C}|}} \approx \sqrt{p}$. The approximation holds as long as $|\mathcal{L}|$, $\mathcal{V}$, and n are not all too small. For example, as shown in Table 2, the percentage of 5-gram words with unique NER labels is 99.37, so ideally the corresponding `PABI` will be $\sqrt{0.9937} = 0.9968$. It is worthwhile to note that the k-gram words with unique labels can also be caused by the low frequency of the appearance of the k-grams. In our experiments, we only consider the k-grams with unique labels that appear at least twice in the data. We experiment on NER with three types of cross-sentence constraints: uni-gram words with unique NER labels, bi-gram words with unique NER labels, and 5-gram part-of-speech (PoS) tags with unique NER labels[8]. The results are shown in Fig. 4.

Auxiliary labels. For auxiliary labels, we show two examples as follows:

- For a multi-class sequence tagging task, we use the corresponding detection task as auxiliary signals. Given a multi-class sequence tagging task with C labels in the BIO format (Ramshaw & Marcus, 1999), we will have 3 labels for the detection and $2C + 1$ labels for the classification. Thus, $\hat{S}(\pi_0, \tilde{\pi}_0) = S(\pi_0, \tilde{\pi}_0) = S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{(1-p_o)\ln C}{\ln(2C+1)}}$, where p_o is the percentage of the label O among all labels.
- Coarse-grained NER for Fine-grained NER. We have four types, PER, ORG, LOC and MISC for CoNLL NER and 18 types for Ontonotes NER. The mapping between CoNLL NER and Ontonotes NER is as follows: PER (PERSON), ORG (ORG), LOC(LOC, FAC, GPE), MISC(NORP, PRODUCT, EVENT, LANGUAGE), O(WORF_OF_ART, LAW, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL, O) (Augenstein et al., 2017). In the BIO setting, we have $\hat{S}(\pi_0, \tilde{\pi}_0) = S(\pi_0, \tilde{\pi}_0) = S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{P_l \ln 3 + P_m \ln 4 + P_o \ln 19}{\ln 37}}$, where p_l, p_m, p_o are the percentage of LOC(including B-LOC and I-LOC), MISC (including B-MISC and I-MISC), and O among all possible labels.

Note that `PABI` is consistent with the entropy normalized mutual information (see more in footnote 5) because $\hat{S}(\pi_0, \tilde{\pi}_0) = \sqrt{\frac{I(Y;\tilde{Y})}{H(Y)}}$ for auxiliary labels.

Cross-lingual signals. For cross-lingual signals, we can use multilingual BERT to get $\hat{c}$ in the extended input space $(\mathcal{V} \cup \mathcal{V}')^n$. After that, η_1 and η_2 can be computed accordingly.

Cross-modal signals. For cross-modal signals, we only consider the case where labels of gold and incidental signals are same and inputs of gold and incidental are aligned. A common situation is that a video has visual, acoustic, and textual information. In this case, the images and speech related to the texts can be used as cross-modal information. We can use cross-modal mapping between speech/images and texts (e.g. Chung et al. (2018)) to estimate the η_1 and η_2 for cross-modal signals.

The mix of cross-domain signals and constraints. Let $\tilde{c}$ denote the perfect system on cross-domain signals and satisfying constrains on inputs of gold signals, and $\hat{c}$ denote the model trained on cross-domain signals and satisfying constraints on inputs of gold signals. In this way, we can estimate η_1 and η_2 by forcing constraints in their inference stage.

A.7 DERIVATION OF EQUATION (4)

For simplicity, we use Y to denote $c(\mathbf{x})$, $\tilde{Y}$ to denote $\tilde{c}(\mathbf{x})$, and $\hat{Y}$ to denote $\hat{Y}(\mathbf{x})$. We then rewrite the definitions of η, η_1' and η_2 as $\eta = \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{D}}(\mathbf{x})} \mathbf{1}(c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})) = P(Y \neq \tilde{Y})$, $\eta_1' = \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{D}}(\mathbf{x})} \mathbf{1}(\hat{c}(\mathbf{x}) \neq \tilde{c}(\mathbf{x})) = P(\hat{Y} \neq \tilde{Y})$ and $\eta_2 = \mathbb{E}_{\mathbf{x} \sim P_{\mathcal{D}}(\mathbf{x})} \mathbf{1}(\hat{c}(\mathbf{x}) \neq c(\mathbf{x})) = P(\hat{Y} \neq Y)$. Note that $\mathcal{L}$ is the label set for the task. Considering all three systems in the target domain, we have

$$\begin{aligned}
1 - \eta_2 &= P(\hat{Y} = Y) \\
&= P(\hat{Y} = Y, \tilde{Y} = Y) + P(\hat{Y} = Y, \tilde{Y} \neq Y) \\
&= P(\tilde{Y} = Y) P(\hat{Y} = Y | \tilde{Y} = Y) + P(\tilde{Y} \neq Y) P(\hat{Y} = Y | \tilde{Y} \neq Y) \\
&= P(\tilde{Y} = Y) P(\hat{Y} = \tilde{Y}) + P(\tilde{Y} \neq Y) \frac{P(\hat{Y} \neq \tilde{Y})}{|\mathcal{L}| - 1} \\
&= (1 - \eta)(1 - \eta_1') + \frac{\eta \eta_1'}{|\mathcal{L}| - 1}
\end{aligned}$$

---

[8] Here we use PoS tags as a special type of cross-sentence constraints by specifying the labels of tokens whose PoS tags have unique NER labels, although PoS tags can also be viewed as auxilary signals for NER.

Therefore, we have $\eta = \frac{(|\mathcal{L}|-1)(\eta_1'-\eta_2)}{1-|\mathcal{L}|(1-\eta_1')}$.

A.8 PABI FOR TRANSDUCTIVE SIGNALS

Assumption I: $\tilde{c}(\mathbf{x})$ is a noisy version of $c(\mathbf{x})$ with a noise ratio η in both target and source domain: $\eta = \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}(\mathbf{x})}\mathbf{1}(c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})) = \mathbb{E}_{\mathbf{x}\sim P_{\tilde{\mathcal{D}}}(\mathbf{x})}\mathbf{1}(c(\mathbf{x}) \neq \tilde{c}(\mathbf{x}))$.

Theorem A.2. *Let $\mathcal{C}$ be a concept class of VC dimension d for binary classification. Let S_+ be a labeled sample of size m generated by drawing βm points (S) from $\mathcal{D}$ according to c and $(1-\beta)m$ points ($\tilde{S}$) from $\tilde{D}$ (the distribution of incidental signals) according to $\tilde{c}$. If $\hat{c}' = \arg\min_{c\in\mathcal{C}} R_{S_+,\frac{1}{2}}(c) = \arg\min_{c\in\mathcal{C}} \frac{1}{2}R_S(c) + \frac{1}{2}R_{\tilde{S}}(c)$ is the empirical joint error minimizer, and $c_T^* = \arg\min_{c\in\mathcal{C}} R_{\mathcal{D}}(c)$ is the target error minimizer, $c^* = \arg\min_{c\in\mathcal{C}} R_{\tilde{\mathcal{D}}}(c) + R_{\mathcal{D}}(c)$ is the joint error minimizer, under assumption I, and assume that $\mathcal{C}$ is expressive enough so that both the target error minimizer and the joint error minimizer can achieve zero errors, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$R_{\mathcal{D}}(\hat{c}') \leq \eta + 4\sqrt{\frac{1}{\beta} + \frac{1}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}}$$

A concept is a function c: $\mathcal{X} \rightarrow \{0,1\}$. The probability according to the distribution $\mathcal{D}$ that a concept c disagrees with a labeling function f (which can also be a concept) is defined as

$$R_{\mathcal{D}}(c,f) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}}[|c(\mathbf{x}) - f(\mathbf{x})|] \tag{5}$$

Note that here $\ell(y, c(x)) = |y - c(x)|$ is the loss function and $R_{\mathcal{D}}(c) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\ell(y, c(\mathbf{x}))]$ where y is the gold label for $\mathbf{x}$. We denote $R_\alpha(c)$ ($\alpha \in [0,1]$) the corresponding weighted combination of true source and target errors, measured with respect to $\tilde{\mathcal{D}}$ and $\mathcal{D}$ as follows:

$$R_\alpha(c) = \alpha R_{\mathcal{D}}(c) + (1-\alpha)R_{\tilde{\mathcal{D}}}(c)$$

Lemma A.3. *Let c be a concept in concept class $\mathcal{C}$. Then*

$$|R_\alpha(c) - R_{\mathcal{D}}(c)| \leq (1-\alpha)(\Lambda + \tau(c))$$

where $\Lambda = R_{\tilde{\mathcal{D}}}(c^*) + R_{\mathcal{D}}(c^*)$, $c^* = \arg\min_{c\in\mathcal{C}} R_{\tilde{\mathcal{D}}}(c) + R_{\mathcal{D}}(c)$, and $\tau(c) = |R_{\tilde{\mathcal{D}}}(c,c^*) - R_{\mathcal{D}}(c,c^*)|$.

Proof.

$$\begin{aligned}
|R_\alpha(c) - R_{\mathcal{D}}(c)| &= (1-\alpha)|R_{\tilde{\mathcal{D}}}(c) - R_{\mathcal{D}}(c)| \\
&\leq (1-\alpha)[|R_{\tilde{\mathcal{D}}}(c) - R_{\tilde{\mathcal{D}}}(c,c^*)| + |R_{\tilde{\mathcal{D}}}(c,c^*) - R_{\mathcal{D}}(c,c^*)| + |R_{\mathcal{D}}(c,c^*) - R_{\mathcal{D}}(c)|] \\
&\leq (1-\alpha)[R_{\tilde{\mathcal{D}}}(c^*) + |R_{\tilde{\mathcal{D}}}(c,c^*) - R_{\mathcal{D}}(c,c^*)| + R_{\mathcal{D}}(c^*)] \\
&= (1-\alpha)(\Lambda + \tau(c))
\end{aligned}$$

Lemma A.4. *For a fixed concept c from $\mathcal{C}$ with VC dimension d, if a random labeled sample (S_+) of size m is generated by drawing βm points (S) from $\mathcal{D}$ and $(1-\beta)m$ points ($\tilde{S}$) from $\tilde{\mathcal{D}}$, and labeling them according to $f_{\mathcal{D}}$ and $f_{\tilde{\mathcal{D}}}$ respectively, then for any $\delta \in (0,1)$ with probability at least $1-\delta$ (over the choice of the samples),*

$$|R_\alpha(c) - R_{S_+,\alpha}(c)| \leq 2\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{4}{\delta}}{m}}$$

where $R_{S_+,\alpha} = \alpha R_S(c) + (1-\alpha)R_{\tilde{S}}(c)$ and e is the natural number.

Proof. Given Lemma 5 in Ben-David et al. (2010), which says for any $\delta \in (0,1)$, with probability $1-\delta$ (over the choice of the samples),

$$P[|R_{S_+,\alpha}(c) - R_\alpha(c)| \geq \epsilon] \leq 2\exp\left(\frac{-2m\epsilon^2}{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\right)$$

According to the Vapnik-Chervonenkis theory (Vapnik & Chervonenkis, 2015), we have with probability $1-\delta$,

$$|R_\alpha(c) - R_{S_+,\alpha}(c)| \le 2\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{4}{\delta}}{m}}$$

This is the standard generalization bound with an adjust term $\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}$ (see more in Chapter 3.3 of Mohri et al. (2018)). $\square$

Proof of Theorem A.2. Let $\alpha = \frac{1}{2}$, then $\hat{c}' = \arg\min R_{S_+,\alpha}(c) = \frac{1}{2}(R_{\tilde{S}}(c) + R_S(c))$

$$\begin{aligned}
R_{\mathcal{D}}(\hat{c}') &\le R_\alpha(\hat{c}') + (1-\alpha)(\Lambda + \tau(\hat{c}')) \quad \text{(Lemma A.3)}\\
&\le R_\alpha(\hat{c}') + (1-\alpha)(\Lambda + |R_{\tilde{\mathcal{D}}}(\hat{c}', c^*) - R_{\mathcal{D}}(\hat{c}', c^*)|) \quad \text{(Definition of } \tau(\hat{c}')\text{)}\\
&\le R_\alpha(\hat{c}') + (1-\alpha)(\Lambda + R_{\tilde{\mathcal{D}}}(\hat{c}') + R_{\tilde{\mathcal{D}}}(c^*) + R_{\mathcal{D}}(\hat{c}') + R_{\mathcal{D}}(c^*))\\
&\le R_\alpha(\hat{c}') + (1-\alpha)(2\Lambda + 2R_\alpha(\hat{c}'))\\
&= (3-2\alpha)R_\alpha(\hat{c}') + 2(1-\alpha)\Lambda\\
&\le (3-2\alpha)(R_{S_+,\alpha}(\hat{c}') + 2\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}})\\
&+ 2(1-\alpha)\Lambda \quad \text{(Lemma A.4 with } \delta/2\text{)}\\
&\le (3-2\alpha)(R_{S_+,\alpha}(c_T^*) + 2\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}})\\
&+ 2(1-\alpha)\Lambda \quad (\hat{c}' = \arg\min R_{S_+,\alpha}(c))\\
&\le (3-2\alpha)(R_\alpha(c_T^*) + 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}})\\
&+ 2(1-\alpha)\Lambda \quad \text{(Lemma A.4 with } \delta/2\text{)}\\
&\le (3-2\alpha)(R_{\mathcal{D}}(c_T^*) + 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}} + (1-\alpha)(\Lambda + \tau(c_T^*)))\\
&+ 2(1-\alpha)\Lambda \quad \text{(Lemma A.3)}\\
&\le (3-2\alpha)(R_{\mathcal{D}}(c_T^*) + 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}})\\
&+ (2\alpha^2 - 7\alpha + 5)\Lambda + (2\alpha^2 - 5\alpha + 3)\tau(c_T^*)
\end{aligned}$$

Note that

$$\tau(c_T^*) = |R_{\tilde{\mathcal{D}}}(c_T^*, c^*) - R_{\mathcal{D}}(c_T^*, c^*)| \le R_{\tilde{\mathcal{D}}}(c_T^*) + R_{\tilde{\mathcal{D}}}(c^*) + R_{\mathcal{D}}(c_T^*) + R_{\mathcal{D}}(c^*) = \Lambda + R_{\mathcal{D}}(c_T^*) + R_{\tilde{\mathcal{D}}}(c_T^*)$$

Therefore,

$$R_{\mathcal{D}}(\hat{c}') \le R_{\tilde{\mathcal{D}}}(c_T^*) + 4\sqrt{\frac{1}{\beta} + \frac{1}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}} + 3R_{\mathcal{D}}(c_T^*) + 3\Lambda \quad (\alpha = \frac{1}{2})$$

Also note that L_1 loss is equivalent to 0-1 loss in the binary classification, so that $R_{\tilde{\mathcal{D}}}(c_T^*) = \eta$ under assumption I. In addition, assuming that $\mathcal{C}$ is expressive enough so that both the target error minimizer and the joint error minimizer can achieve zero errors ($R_{\mathcal{D}}(c_T^*) = 0$ and $\Lambda = 0$), the generalization bound can be simplified as follows:

$$R_{\mathcal{D}}(\hat{c}') \le \eta + 4\sqrt{\frac{1}{\beta} + \frac{1}{1-\beta}}\sqrt{\frac{2d\ln\frac{2em}{d} + 2\ln\frac{8}{\delta}}{m}} \quad \square$$

Note that the proof of Theorem A.2 is similar to Theorem 3 in Ben-David et al. (2010). Our theorem is based on binary classification mainly because the error item in Eq. (5) based on the L1 loss will be equivalent to zero-one loss for binary classification. Although for multi-class classification, the L1 loss is different from commonly used zero-one loss, Theorem A.2 also indicates the relation between the generalization bound of joint training and the cross-domain performance $R_{\mathcal{D}}(c_S^*)$ (equal to $R_{\tilde{\mathcal{D}}}(c_T^*)$ under assumption I). Furthermore, a multi-class classification task can be represented by a series of binary classification tasks. Therefore, we postpone more accurate analysis for multi-class classification as our future work.

A.9 DETAILS OF EXPERIMENTAL SETTINGS

In this subsection, we briefly highlight some important settings in our experiments and more details can be found in our submitted code.

NER with individual inductive signals. For partial labels, we experiment on NER with four different partial rates: 0.2, 0.4, 0.6, and 0.8. For noisy labels, we experiment on NER with seven different noisy rates: $0.1 - 0.7$. For auxiliary labels, we experiment on two auxiliary tasks: named entity detection and coarse NER (CoNLL annotations with 4 types of named entities (Sang & De Meulder, 2003)).

NER with mixed inductive signals. A more complex case is the comparison between the mixed inductive signals. For the first type of mixed signals, we experiment on the combination between three unknown partial rates (0.2, 0.4, and 0.6) and four noisy rates (0.1, 0.2, 0.3, and 0.4). As for the second type of mixed signals, we experiment on the combination between the BIO constraint and five unknown partial rates (0.2, 0.4, 0.6, 0.8, and 1.0).

NER with various inductive signals. After we put the three types of individual inductive signals and the two types of mixed inductive signals together, we still see a correlation between `PABI` and the relative performance improvement in experiments in Fig. 2(f).

NER with cross-domain signals Because we only focus on the person names, a lot of sentences in the original dataset will not include any entities. We random sample sentences to keep that 50% sentences without entities and 50% sentences with at least one entity. η_1 and η_2 is computed by using sentence-level accuracy.

QA with cross-domain signals. For consistency, we only keep one answer for each question in all datasets. Another thing worthwhile to notice is that the most informative QA dataset is not always the same for different main QA datasets. For example, for NewsQA, the most informative QA dataset is SQuAD, while the most informative QA dataset for SQuAD is QAMR.

Experimental settings for learning with various inductive signals. The 2-layer NNs we use in CWBPP (algorithm 1) has a hidden size of 4096, ReLU non-linear activation and cross-entropy loss. As for the embeddings, we use 300 dimensional Glove embeddings (Pennington et al., 2014). The size of the training batch is 10000 and the optimizer is Adam (Kingma & Ba, 2015) with learning rate $3e^{-4}$. When we initialize the classifier with gold signals (line 1), the number of training epochs is 20. After that, we conduct the bootstrapping 5 iterations (line 3-7). The confidence for predicted labels is exactly the predicted probability of the classifier (line 5). In each iteration of bootstrapping, we further train the classifier on the joint data 1 epoch (line 7).

Experimental settings for learning with cross-domain signals. As for BERT, we use the pre-trained case-insensitive BERT-base pytorch implementation (Wolf et al., 2019). We use the common parameter settings for our experiments. Specifically, for NER, the max length is 256, batch size is 8, the epoch number is 4 and the learning rate is $5e^{-5}$. As for QA, the max length is 384, bath size is 16, the epoch number is 4, and the learning rate is $5e^{-5}$.

A.10 THE CWBPP ALGORITHM

The CWBPP algorithm is shown in Algorithm 1.

Algorithm 1: Confidence-Weighted Bootstrapping with Prior Probability. The algorithm utilizes incidental signals to improve the inference stage in semi-supervised learning.

```
Input: A small dataset with gold signals $\mathcal{D} = (X_1, Y_1)$, and a large dataset with inductive signals
       $\tilde{\mathcal{D}} = (X_2, \tilde{Y}_2)$ where $X_1 \cap X_2 = \phi$
Initialize claissifier $\hat{c} = \textsc{Learn}(\mathcal{D})$ (initialize the classifier with gold signals)
$P(Y_2|X_2, \tilde{Y}_2)$ = $\textsc{Prior}(\mathcal{D}, \tilde{\mathcal{D}})$ (estimate the probability of gold labels for inputs in $\tilde{D}$)
while convergence criteria not satisfied do
    $\hat{Y} = \textsc{Inference}(X_2; \hat{c}; P(Y_2|X_2, \tilde{Y}_2))$ (get predicted labels of inputs in $\tilde{D}$)
    $\hat{\rho} = \textsc{Confidence}(X_2; \hat{c}, P(Y_2|X_2, \tilde{Y}_2))$ (get confidence for predicted labels)
    $\tilde{D} = (X_2, \hat{Y}, \hat{\rho})$ (get confidence-weighted incidental dataset with predicted labels)
    $\hat{c} = \textsc{Learn}(\mathcal{D} + \tilde{\mathcal{D}})$ (learn a classifier with both gold dataset and incidental dataset)
return $\hat{c}$
```