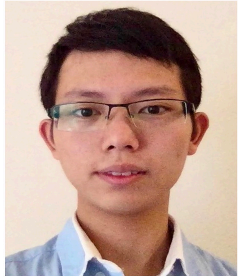
Foreseeing the Benefits of Incidental Supervision



Hangfeng He[†], Mingyuan Zhang[†], Qiang Ning[‡*], and Dan Roth[†]

[†]University of Pennsylvania

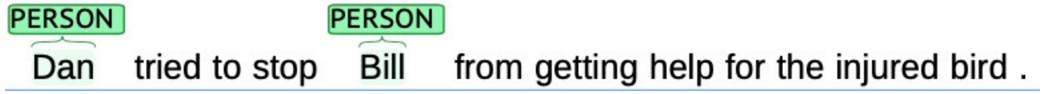
[‡]Amazon

*Part of this work was done while the author was at Allen Institute for AI.

EMNLP, 2021

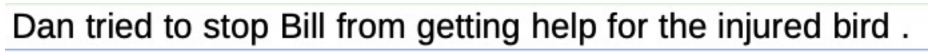
Incidental Supervision Signals

- Given the task of NER, what types of signals can we use?



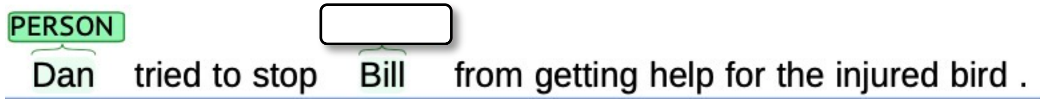
Dan tried to stop Bill from getting help for the injured bird .

Gold Annotations



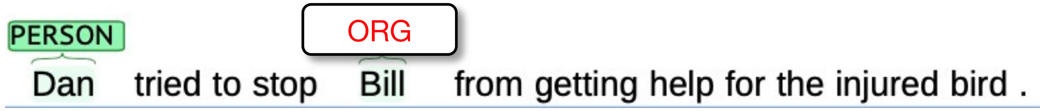
Dan tried to stop Bill from getting help for the injured bird .

Unlabeled texts



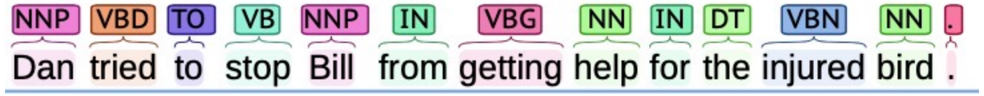
Dan tried to stop Bill from getting help for the injured bird .

Partial Annotations



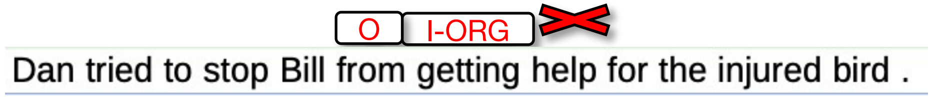
Dan tried to stop Bill from getting help for the injured bird .

Noisy Annotations



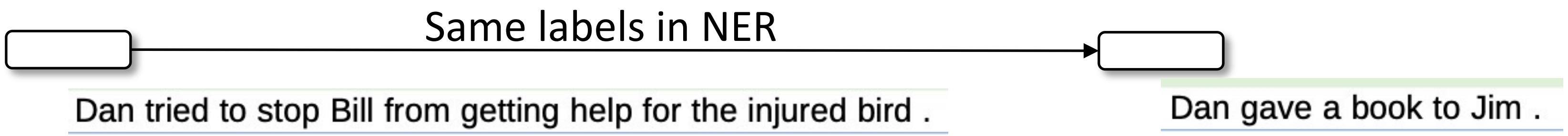
Dan tried to stop Bill from getting help for the injured bird .

Auxiliary Annotations



Dan tried to stop Bill from getting help for the injured bird .

Constraints



Dan tried to stop Bill from getting help for the injured bird .

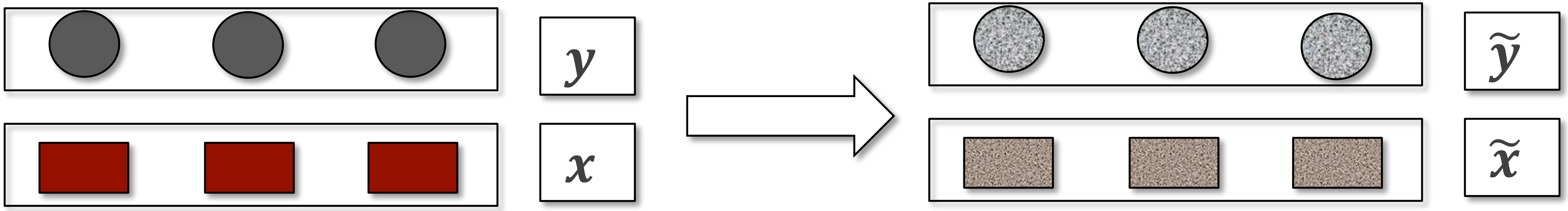
Dan gave a book to Jim .

Knowledge

傅達仁$_{\mathrm{PERSON}}$今將執行安樂死，卻突然爆出自己20年前$_{\mathrm{DATE}}$遭緯來體育台$_{\mathrm{ORG}}$封殺，他不懂自己哪裡得罪到電視台。

Cross-lingual Annotations

Incidental Supervision [Roth, AAAI’17]



Supervised Signals

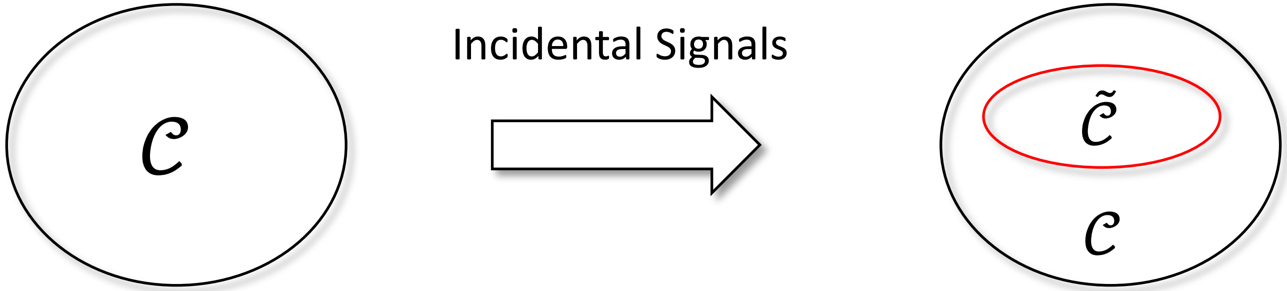
Incidental Signals

- Incidental signals
  - □ Inductive Incidental Signals
    - Partial labels, noisy labels, auxiliary labels, constraints, etc.
  - □ Transductive Incidental Signals
    - Cross-domain signals, cross-lingual signals, cross-modal signals, etc.
  - □ Mixed Incidental Signals
    - Partial + noisy, partial + constraints, cross-domain + noisy, etc.

Can we provide a unified framework for incidental signals, and quantify the extent to which various incidental signals can help the target task?

The Impact of Incidental Signals on the Concept Class

- $c: X \rightarrow Y, where\ c \in \mathcal{C}$
- Learning theory shows that the size of the concept class determines the “easiness” of the learning problem
  - □ E.g. the generalization bound $R(c) \leq \hat{R}(c) + \sqrt{\frac{\ln|\mathcal{C}|+\ln\frac{2}{\delta}}{2m}}$
- We will show that the use of incidental signals reduces the size of the concept class, and then will use the relative size of the reduction as a measure for the informativeness of the incidental signals



Incidental Signals

Original Concept Class

Reduced Concept Class

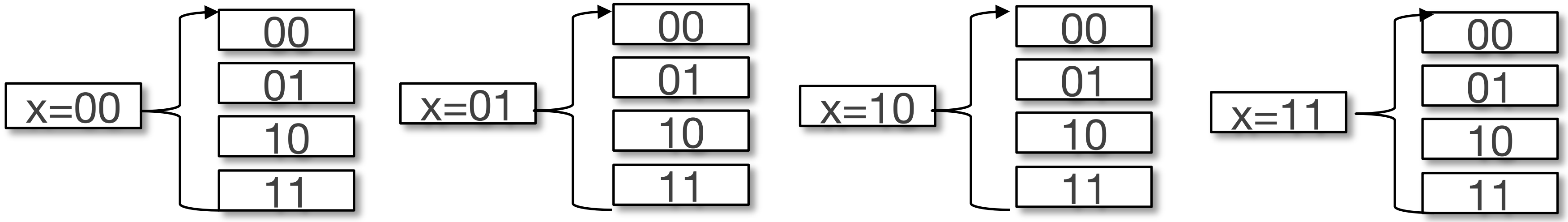
$$S(\mathcal{C}, \tilde{\mathcal{C}}) = \sqrt{1 - \frac{\ln|\tilde{\mathcal{C}}|}{\ln|\mathcal{C}|}}$$

Smaller $\tilde{\mathcal{C}}$ leads to higher Informativeness S

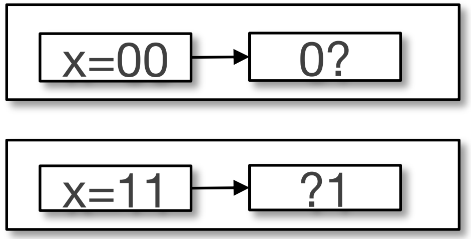
Reduce the concept class from $\mathcal{C}\ \ to\ \tilde{\mathcal{C}}$

An Example with Partial Data

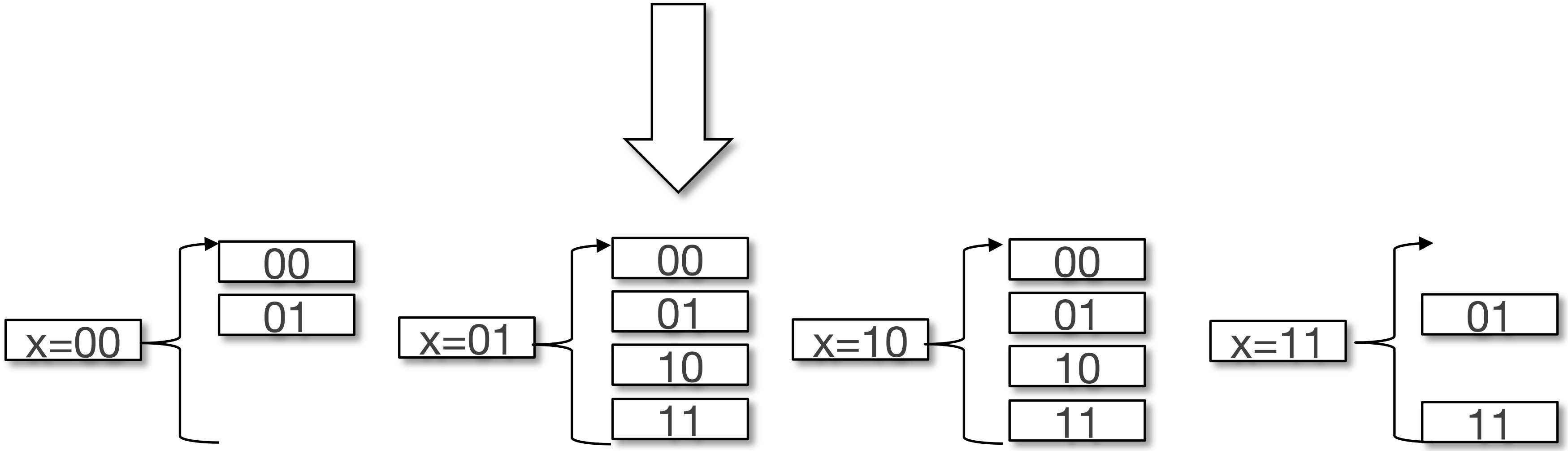
- c: $X \rightarrow Y$, with $X = \{0,1\}^2$ and $Y = \{0,1\}^2$



$|\mathcal{C}| = 4^4 = 256$

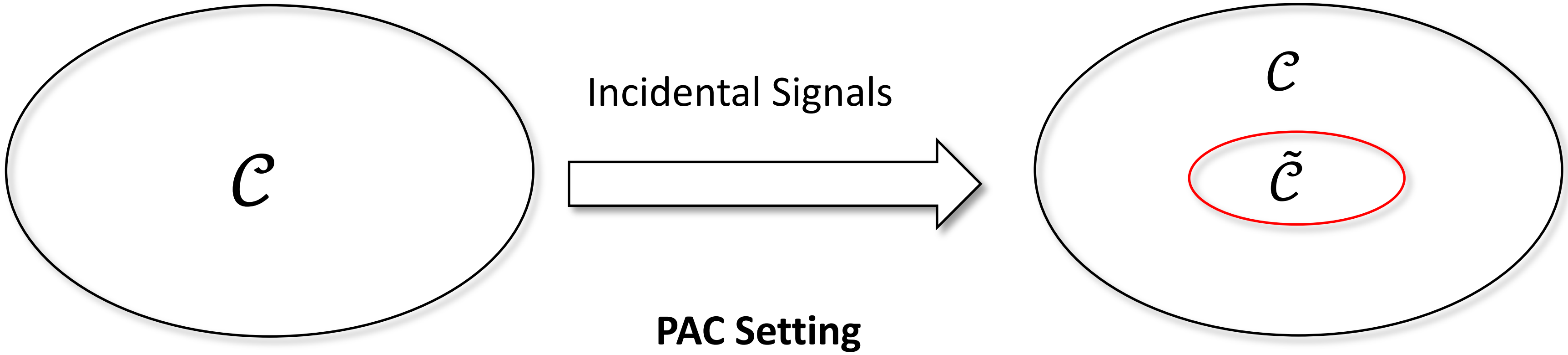


Partial dataset



$| \tilde{\mathcal{C}} | = 2 * 4 * 4 \; * 2 = 64$

PABI: A Unified PAC-Bayesian Informativeness Measure

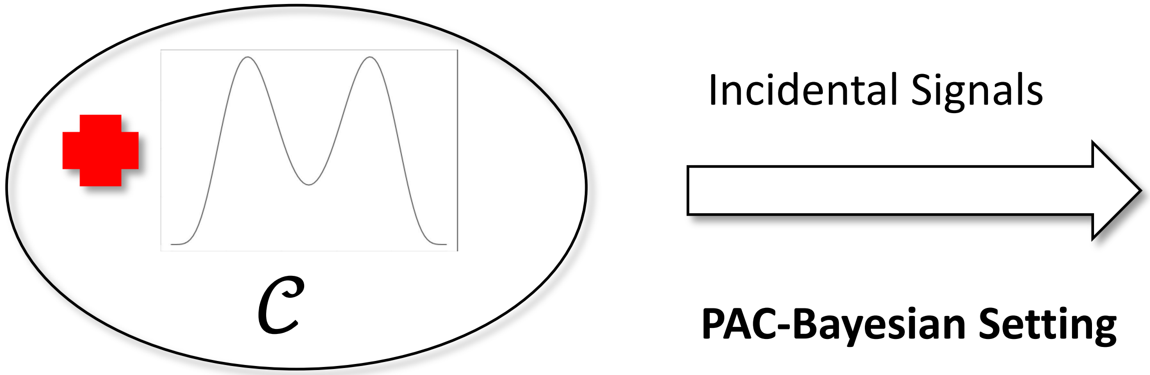


Original Concept Class

Reduced Concept Class

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

Reduce the concept class from $\mathcal{C}$ *to* $\tilde{\mathcal{C}}$



Concept Class with Probability Measure

$$S'(\pi_0,\tilde{\pi}_0) = \sqrt{1 - \frac{D_{KL}(\pi^*||\tilde{\pi}_0)}{D_{KL}(\pi^*||\pi_0)}} \approx \hat{S}'(\pi_0,\tilde{\pi}_0) = \sqrt{1 - \frac{H(\tilde{\pi}_0)}{H(\pi_0)}}$$

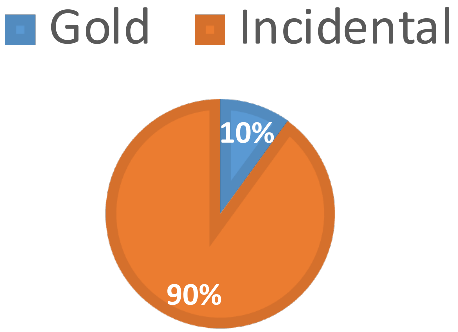
Make the prior π_0 closer to the gold posterior π^*

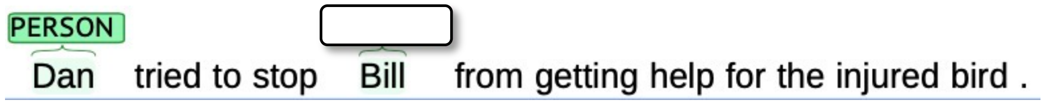
Can handle the infinite concept class case

For non-probabilistic cases, $S = S' = \hat{S}'$

Learning with Various Inductive Incidental Signals

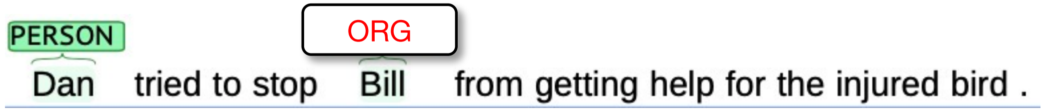
- Main Task
  - Named entity recognition (NER)
- Data
  - Ontonotes 5.0 (18 types of named entities)
- Basic Model
  - Two-layer NNs with 5-gram features
- Algorithm
  - Bootstrapping with incidental signals
    - Initialize the model by training it on the gold signals
    - An EM algorithm on the large-scale incidental signals
    - **Improve the inference stage with incidental signals**
- Setting
  - 200K word-level examples





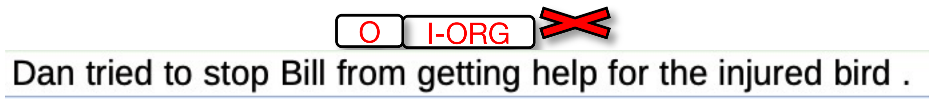
PERSON
Dan tried to stop Bill from getting help for the injured bird .

Partial Annotations



PERSON
ORG
Dan tried to stop Bill from getting help for the injured bird .

Noisy Annotations



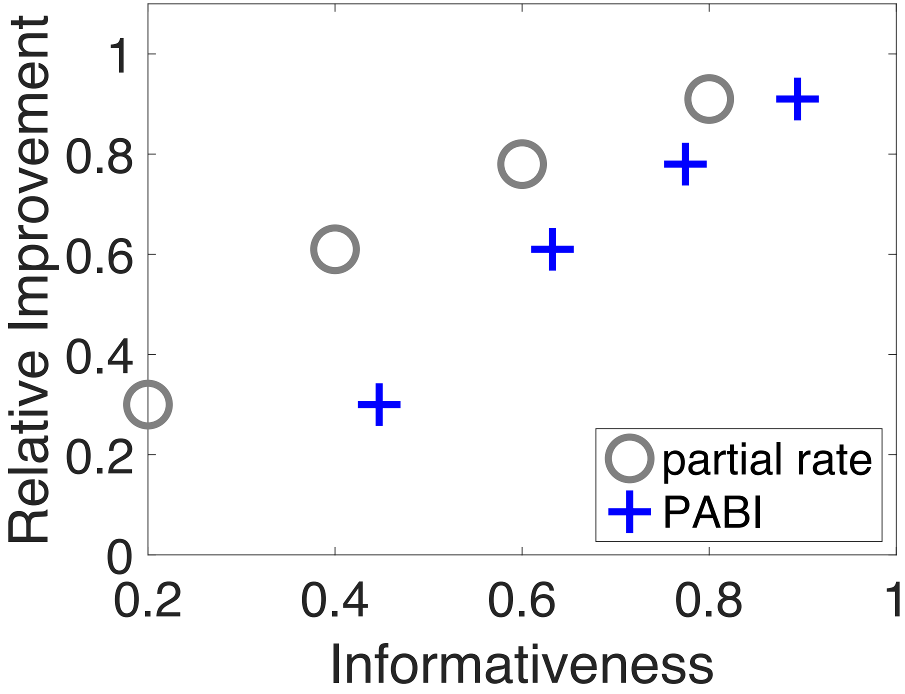
O
I-ORG
Dan tried to stop Bill from getting help for the injured bird .

Constraints

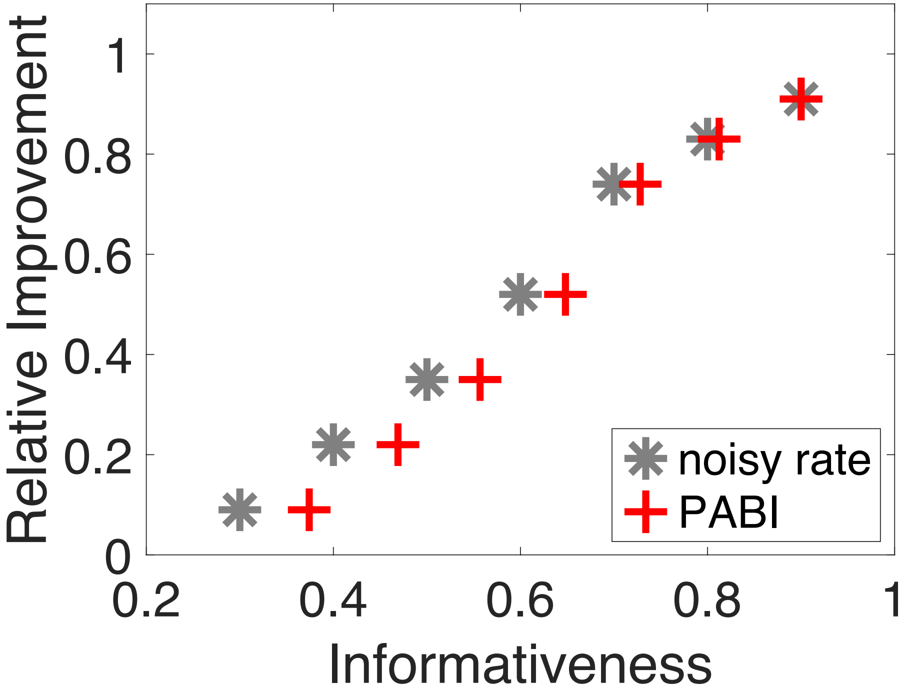
What are our expectations when we add various incidental signals to small amounts of gold annotations?

Can we determine, ahead of time, the relative benefits of partial data, noisy data and constraints?

Results: Individual Inductive Signals



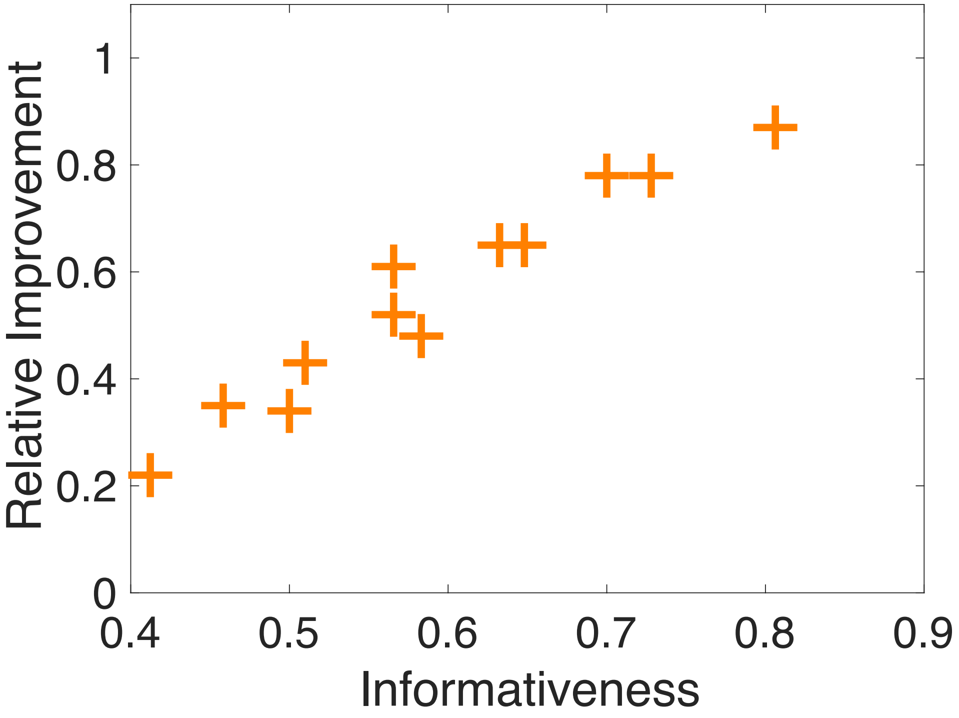
Partial supervision: the relation between the relative improvement and the informativeness for partial signals with different unknown partial rates



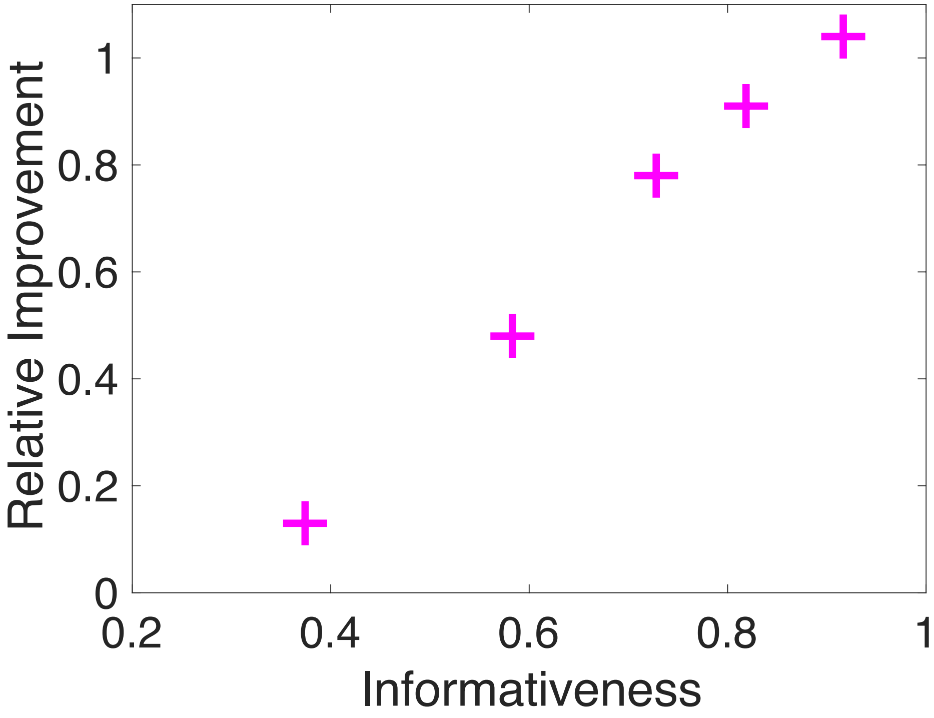
Noisy supervision: the relation between the relative improvement and the informativeness for noisy signals with different noise rates

It is not surprising that:
(1) Signals with lower unknown partial rates lead to higher improvement
(2) Signals with lower noise rates lead to higher improvement

Results: Mixed Inductive Signals



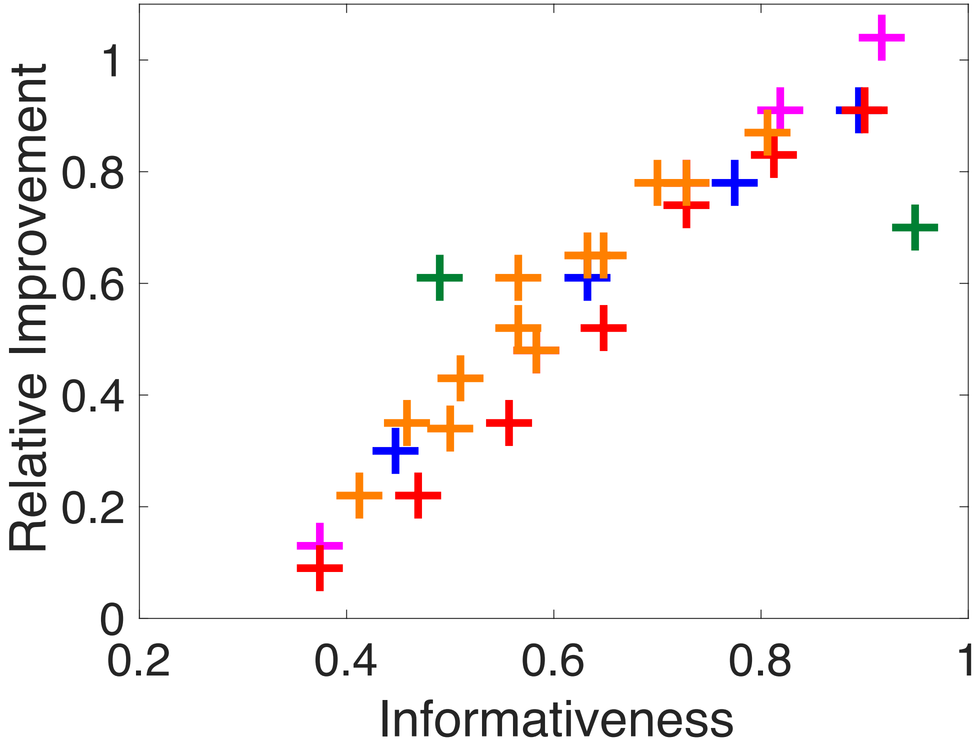
Partial + noisy supervision: the relation between the relative improvement and the informativeness for signals with both partial and noisy signals



Partial + constraints supervision: the relation between the relative improvement and the informativeness for signals with both partial labels and constraints

The (relative) benefits from the mixed signals (e.g., a dataset is both partial and noisy) cannot be determined in existing frameworks, but our framework can handle it.

Results: Various Inductive Signals



The relation between the relative improvement and PABI for various incidental signals: partial labels, noisy labels, auxiliary labels, partial + noisy, and partial + constraints.

The Pearson’s correlation coefficient is: 0.92

The Spearman's rank correlation coefficient is: 0.93

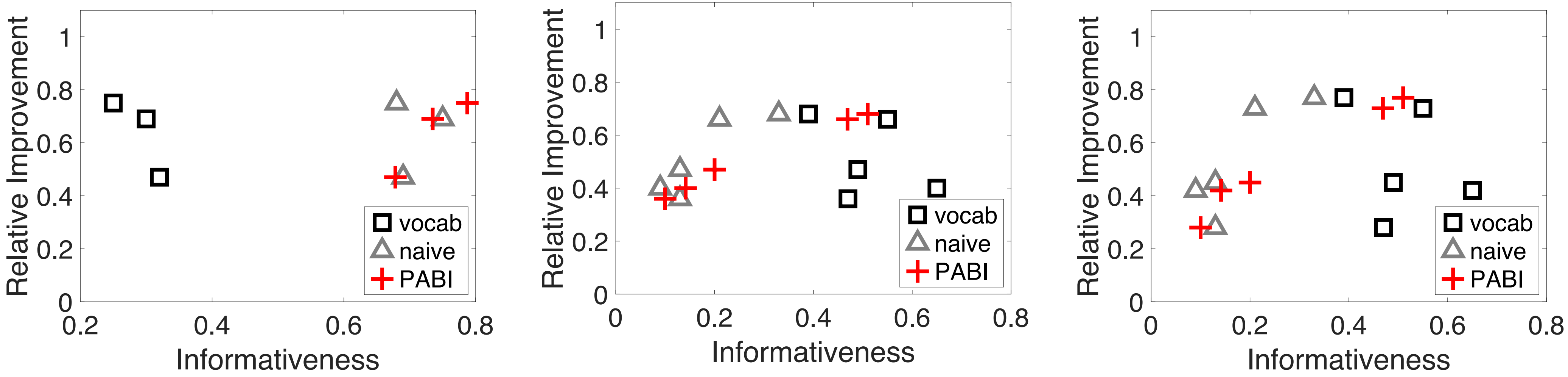
Take away:
The informativeness of a signal predicts the improvement provided by the signal.

Key Insight:
It is possible to compare (and predict) the potential contribution of incidental supervision of different types with PABI.

Experiments: Cross-Domain Signals

- NER datasets (person entities)
  - Main dataset (85 sentences): twitter
  - Cross-domain datasets (756 sentences): Ontonotes, CoNLL, and GMB
- QA datasets
  - Main dataset (700 QA paris): SQuAD
  - Cross-domain datasets (6.2 K QA pairs): QAMR, QA-SRL Bank 2.0, QA-RE, NewsQA, Trivia QA
- Model
  - BERT
  - Joint training / pre-training
- Informativeness Measures
  - **Vocabulary overlap baseline** (Gururangan et al., 2020)
  - **Naive baseline**: the error rate of the system trained on the cross-domain signals when evaluated on the target domain
  - **PABI:** $\hat{S}'(\pi_0, \tilde{\pi}_0) = \sqrt{1 - \frac{\eta \ln(|\mathcal{L}|-1) - \eta \ln \eta - (1-\eta) \ln(1-\eta)}{\ln|\mathcal{L}|}}$

Results: Cross-Domain Signals



(a) Joint-training NER with baselines and PABI

(b) Joint-training QA with baselines and PABI

(c) Pre-training QA with baselines and PABI

Correlation between informativeness measures (baselines or the PABI) and relative improvement (via joint training or pre-training) for cross-domain NER and cross-domain QA.

The Pearson's correlation of three informativeness measures (PABI, naïve baseline, vocabulary overlap baseline) in the three cases are: 0.96/0.19/-0.85, 1.00/0.88/-0.40, 0.99/0.85/-0.30.

Discussions

- Contributions
  - □ PABI: A **unified** PAC-Bayesian informativeness measure for incidental signals
  - □ Experiments on NER and QA support **the effectiveness** of PABI on quantifying the value of **a wide range of incidental signals**
    - Partial labels, noisy labels, constraints, auxiliary signals, cross-domain signals, and some combinations of them
- Discussion of Factors in PABI
  - □ Base model performance (the size of gold signals)
  - □ The size of incidental signals
  - □ Data distribution
  - □ Algorithm
  - □ Cost-sensitive loss