Data-Dependent Annotator Accuracy for Active Learning

Nathan Hatch and Erik Wijmans

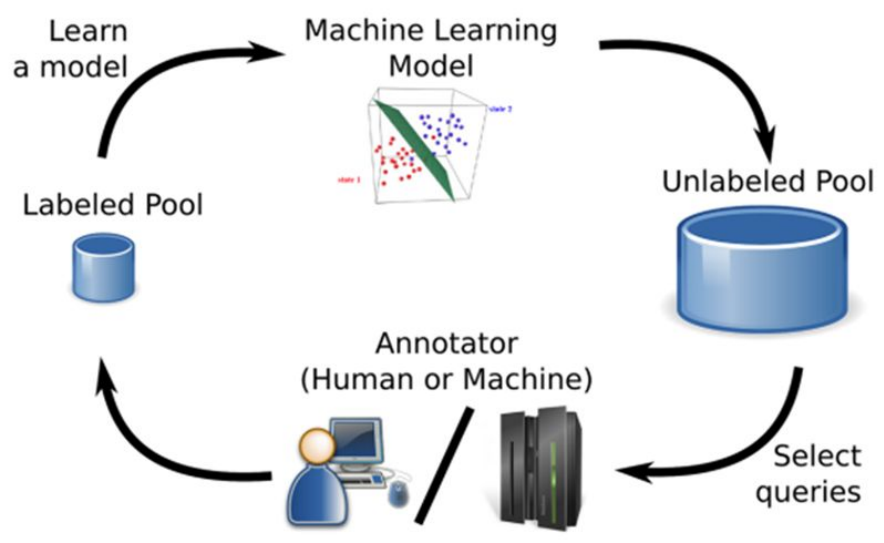
Outline

- **Introduction**
- Methodology
- Experimental Setup
- Results

Why Active Learning

Collecting labeled training examples is expensive.

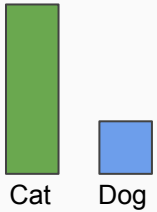
Idea: Have the ML model ***choose*** the most important examples to label.



Uncertainty Sampling

How to choose which examples to label?

Idea: Pick the examples that the model is most ***uncertain*** about.
(e.g. maximum entropy over possible labels)



Challenge: Noisy Annotators

Early active learning researchers assumed ***perfect annotation***.

In practice, annotators are often ***noisy*** (e.g. Amazon Mechanical Turk).

To account for this, active learning algorithms must be able to choose an annotator for each sample.

Outline

- Introduction
- **Methodology**
- Experimental Setup
- Results

Previous approach: IEThresh (Donmez et al. 2009)

Donmez et al. (2009) address this with ***IEThresh***:

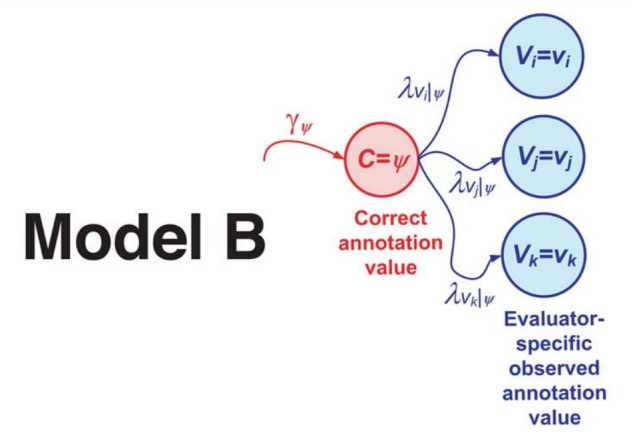
- Estimate individual accuracy based on percent agreement with majority
- Choose the annotator(s) with highest Upper Confidence Interval

This does not account for the possibility that annotator accuracy ***depends on the latent class***.

Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268. ACM, 2009.

Previous approach: Data-Dependent Annotator Accuracy

Rzhetsky et al. (2009) propose a ***partially observed Bayesian network*** to model data-dependent annotator accuracy:



They do not apply this model to the active learning setting.

Andrey Rzhetsky, Hagit Shatkay, and W John Wilbur. How to get the most out of your curation effort. *PLoS computational biology*, 5(5):e1000391, 2009.

Our approach

Use Model B for active learning.

1) Use a “warmup” set of samples to estimate the parameters of Model B.
2) Iteratively:
    a) Choose a batch of samples with uncertainty sampling.
    b) Label each sample using the best annotator according to Model B.

Learning the Bayes Net Parameters

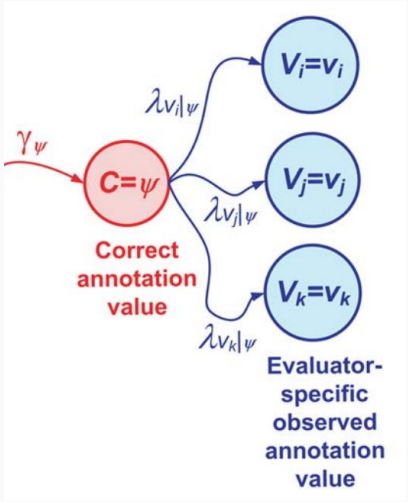
For the warmup set, we obtain labels for each example from ***all*** annotators.

Use Expectation Maximization on this set of ***partially observed samples***.

Initialization matters: Assume annotators are 80% accurate.

True label

?	?	?	?	?



Emitted label

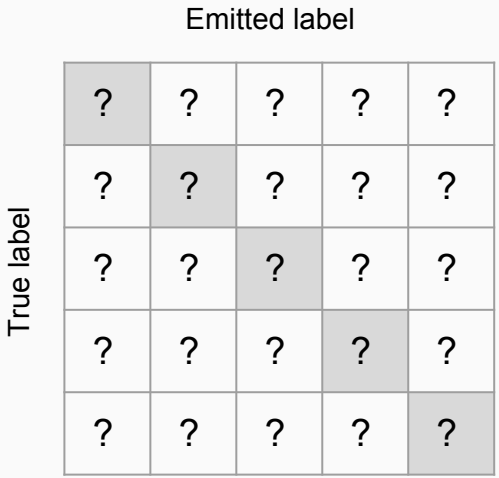
True label

?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

Choosing the Best Annotator

Each unlabeled sample has a predicted class distribution *p*.

Choose the annotator with the highest ***expected accuracy*** for this sample.



Outline

- Introduction
- Methodology
- **Experimental Setup**
- Results

Sentiment Analysis

- Goal to predict sentiment (how positive/negative)
- ***Fully supervised*** dataset
  - Rotten Tomatoes movie reviews
- Sentiment is predicted at phrase level

Sentiment Analysis

- “Deadly dull” has a highly negative sentiment
- “consider a DVD rental” has a neutral sentiment
- “exquisite acting” has a highly positive sentiment

Fake Annotators

- Five fake annotators
- Error rates are ***class specific***
- Three different configurations: Good, Mediocre, Bad

Naive Bayes

- Naive Bayes
  - Focus on annotators/active learning
- Dirichlet prior

Annotator Learning Experiments

- ***How much data*** to accurately model the annotators?
- How do the ***annotator accuracies*** affect our model?

Active Learning Experiments

- How large does the ***warm-up pool*** need to be?
- ***Which annotator*** to give it to?

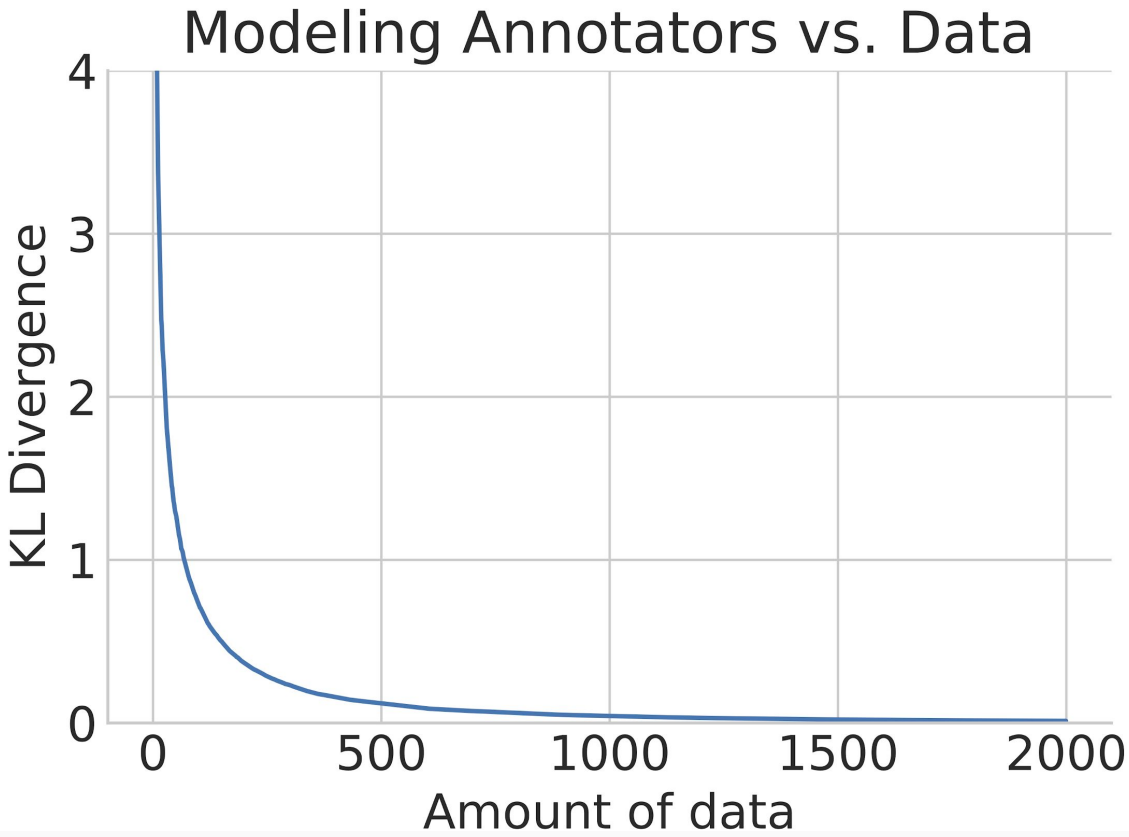
Outline

- Introduction
- Methodology
- Experimental Setup
- **Results**

Modeling Annotators

- **KL divergence** between estimated and true annotator accuracies
- When does the model find the **best annotator** for each class

Modeling Annotators: Accuracy

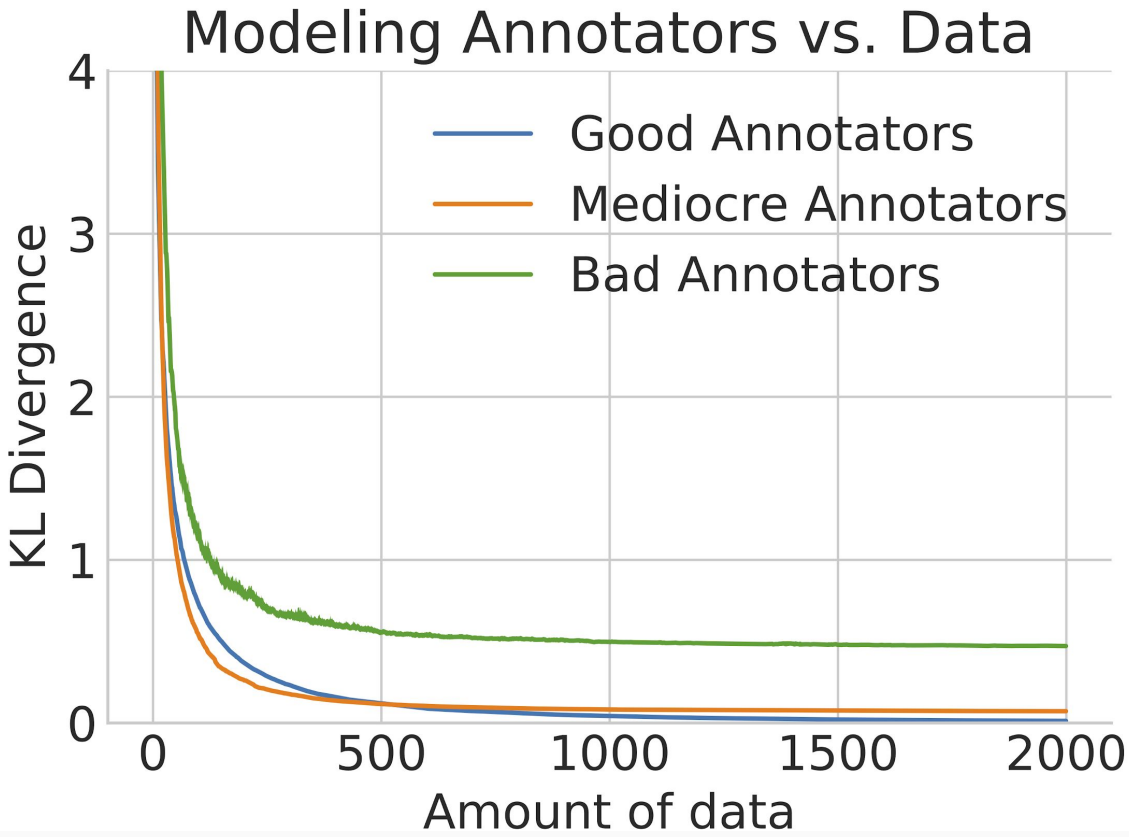


507 5-way annotations

Modeling Annotators

- Three different groups of annotators
  - Good: 0.95 - 0.75
  - Mediocre: 0.9 - 0.5
  - Bad: 0.6 - 0.2
- Same metrics as before

Modeling Annotators: Effect of Annotator Accuracy



507 5-way annotations

324 5-way annotations

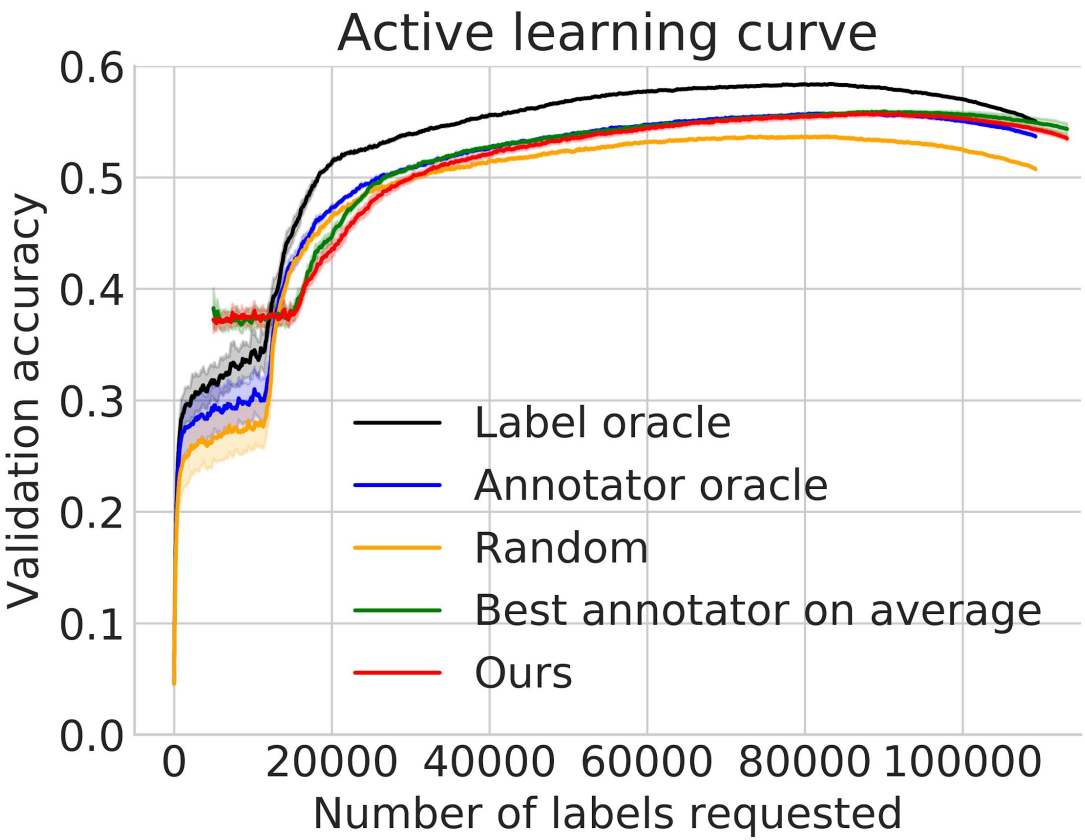
433 5-way annotations

Active Learning Results

Our experiments investigated the following:

- Classifier accuracy vs. number of queries
- Classifier accuracy vs. annotator accuracy
- Classifier accuracy vs. size of warm-up pool

Active Learning Results



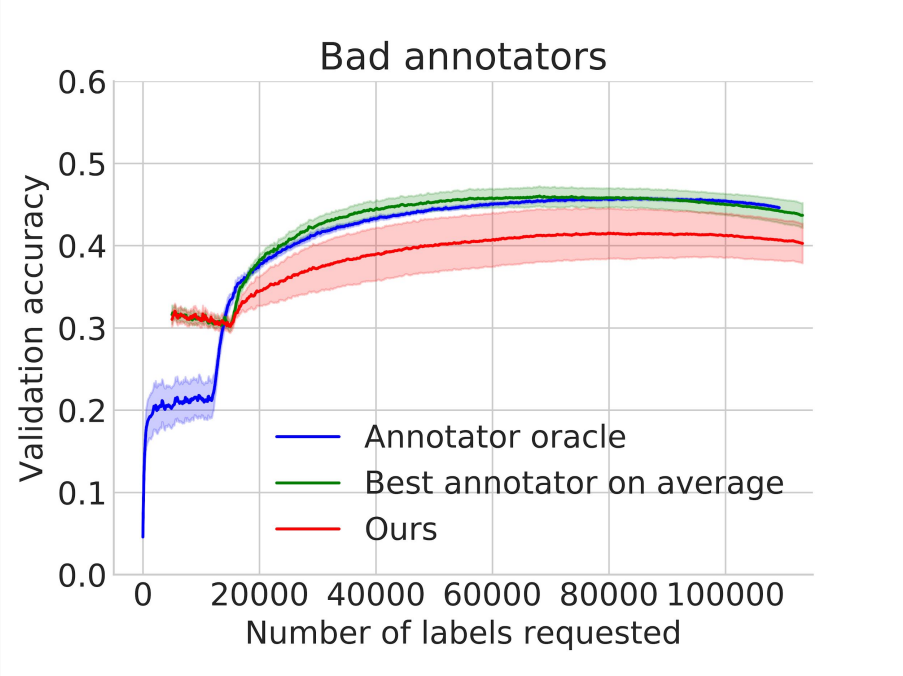
Our method performs as if we knew the true annotator accuracies.

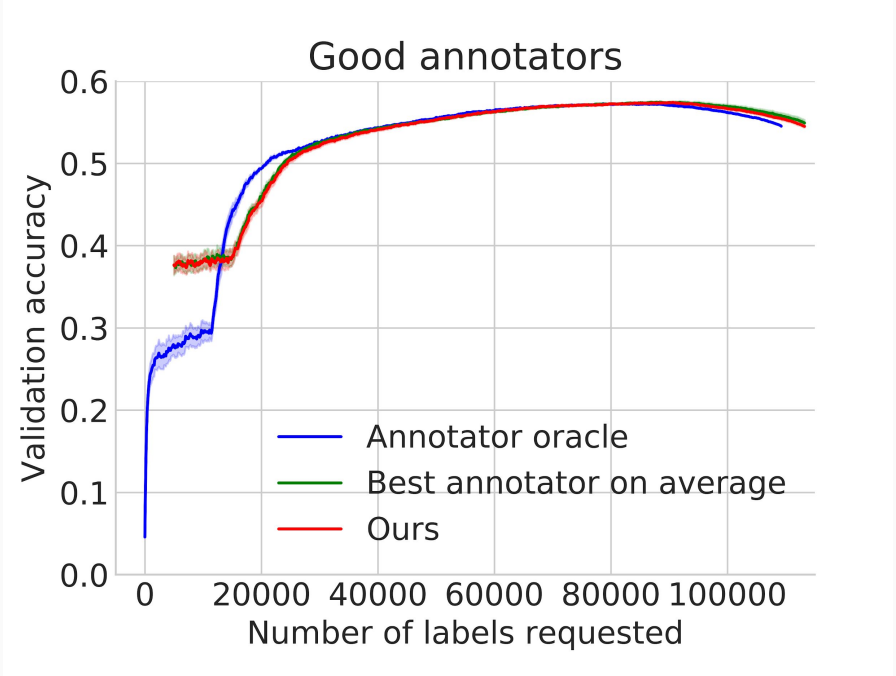
Choosing the “best annotator on average” performs better for this dataset.

Uncertainty sampling has an unusual learning curve for this dataset.

Active Learning Results: Effect of Annotator Accuracy

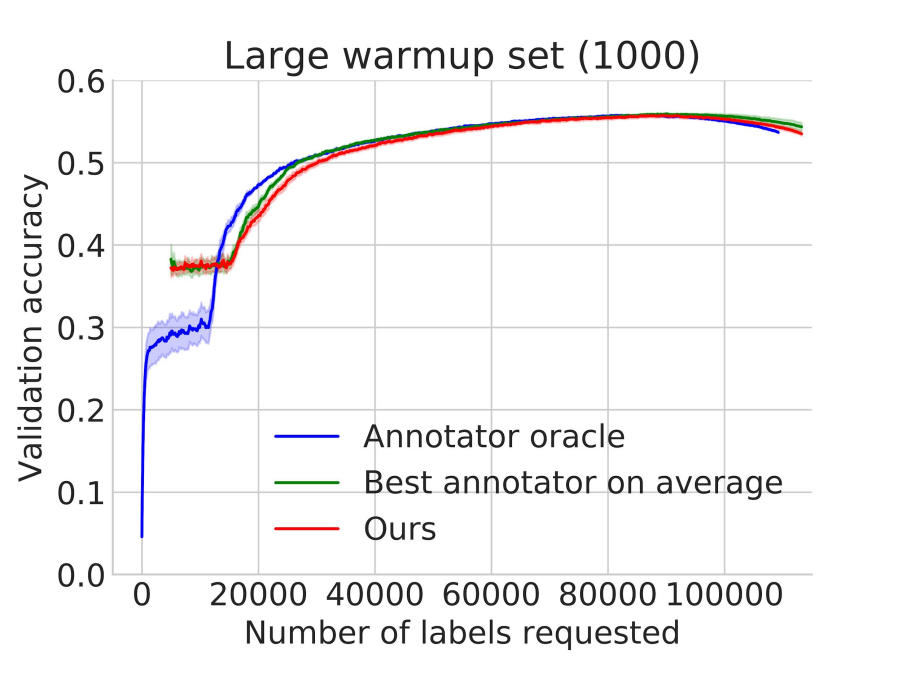
For very accurate annotators, all methods perform about the same.

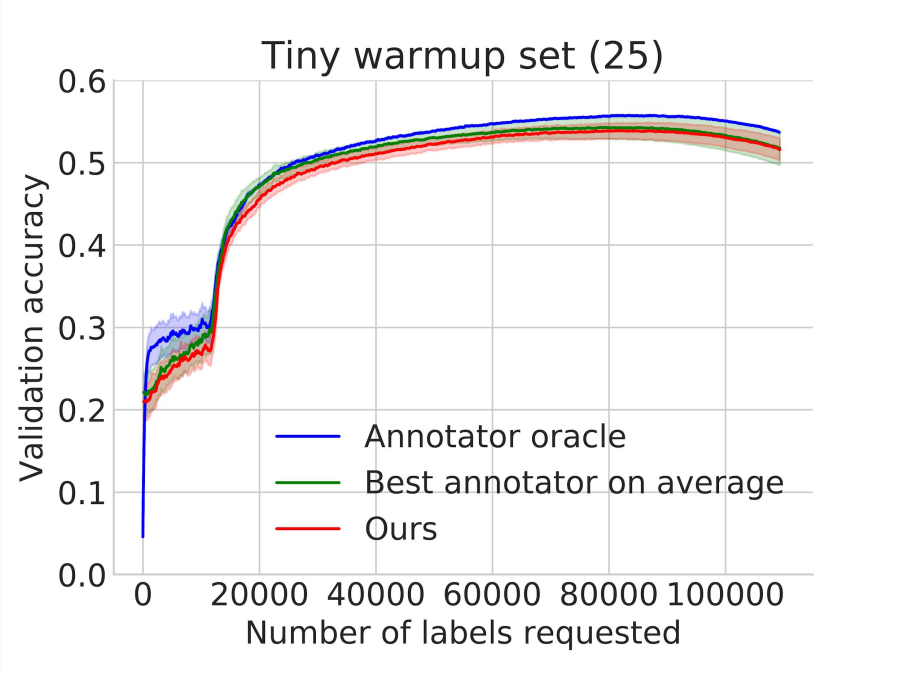




Active Learning Results: Effect of Warm-Up Size

Even for very small warmup sets, all methods perform surprisingly well.





Conclusion and Future Work

Conclusions:

- Modeling annotators works!
- Annotator specific active learning doesn’t seem to help

Future work:

- Use a more expressive hypothesis class than Naive Bayes
- Modify uncertainty sampling to avoid rare words
- Investigate dataset imbalance