Interactive Machine Teaching by Labeling Rules and Instances

COLUMBIA UNIVERSITY

Giannis Karamanolakis
Amazon AGI
karamai@amazon.com

Daniel Hsu
Columbia University
djhsu@cs.columbia.edu

Luis Gravano
Columbia University
gravano@cs.columbia.edu

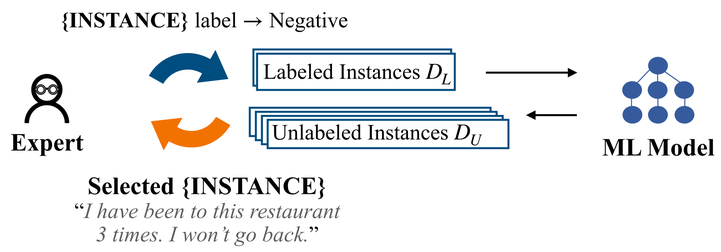
TACL 2024

EMNLP 2025 Suzhou, China 中国苏州

Data labeling is expensive

Approach 1: Active Learning

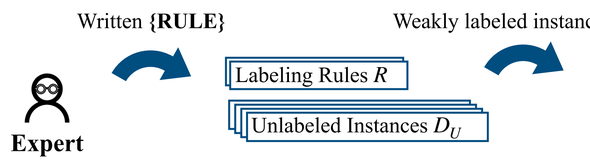
Adaptively choose which instances to label



Issue: Labeling one instance at a time is not scalable.

Approach 2: Weak Supervision

Collect expert-designed rules that automatically create many weak labels



Snorkel (Ratner et al., 2019)
FlyingSquid (Fu et al., 2021)
ASTRA (Karamanolakis et al., 2021)

Rule examples:

Spam classification

```
def regex_check_out(x):
    return SPAM if re.search("check.*out", x) else ABSTAIN
```

Question type classification

```
def numeric_question(x):
    return NUMERIC if x.startswith("when") else ABSTAIN
```

Issue: Designing many strong rules in a single shot is hard.

How to efficiently use expert feedback?

Assume limited budget T to query an expert

- Should we spend T querying for **rules** or **instance labels**?
- What **rule properties** are required to train an accurate model?

Our Work

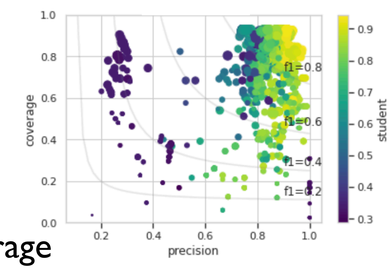
1. Characterization of patterns in Weak Supervision

We unify weak supervision methods using a Teacher-Student abstraction:

- Teacher: uses rules to weakly label D_U
- Student: trained with soft Teacher labels

We evaluate 1,000+ Teacher-Student pairs:

- Higher Teacher $F1 \neq$ higher Student $F1$
- Teacher precision is more important than coverage

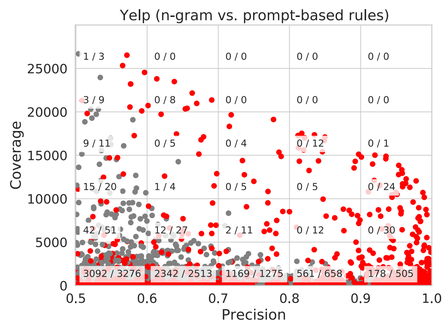


2. Automatic rule extraction via prompting

We propose a method that extracts rules with rich predicates:

- n-gram features
- Syntactic features
- Prompt-based features

Candidate Rules (predicate → label)
PMT-EXPERIENCE="terrible" → Negative
PMT-EXPERIENCE="fantastic"→ Positive
PMT-RECOMMEND="certainly" → Positive
PMT-IS_ABOUT="prizes" → Spam
NGRAM="http" **AND** PMT-ASKS_FOR="donations" → Spam
NER="CARDINAL" **AND** PMT-ASKS_FOR="information"→ Spam



Our rule family achieves higher precision and coverage than n-gram rules.

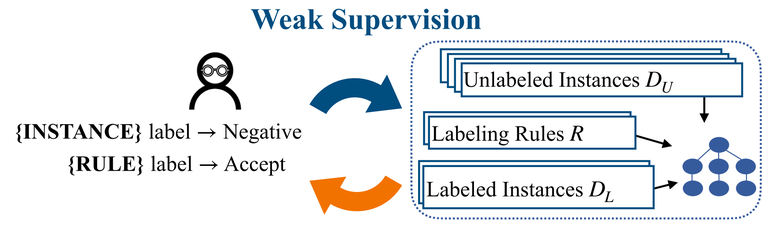
3. Interactive machine teaching

We present a human-in-the-loop machine teaching framework, which queries for expert feedback on both instances and rules.

INTERVAL: Interactive Learning with Weak Supervision

INTERVAL balances the quality of instance labels with the efficiency of labeling rules

Asks expert feedback on both instances and rules



Automatically extracts labeling rules

Selected {INSTANCE}: "I have been to this restaurant 3 times. I won't go back."

Extracted {RULE}: LLM("{INSTANCE}. Overall, the experience is {MASK}") = "terrible"

Algorithm

Initialize $D'_L = D_L$, $R' = R$
Repeat until the budget T runs out:
3.1: Train Teacher $q^*_\phi(\cdot)$ and Student $p_\theta(\cdot)$ using D'_L, D_U, R'
3.2: Apply $p_\theta(\cdot)$ to $s \in D_U$ to obtain soft labels: $D_{Student} = \{(s_i, \mathbf{p}_i)\}_{s_i \in D_U}$
3.3: Pick a candidate instance $s_i \in D_U$
3.4: Query the label y_i for s_i (cost $= T_I$)
3.5: Extract candidate rules r^j that cover s_i
3.6: Query the labels z^j for β_i rules r^j (cost $= \beta_i \cdot T_R$)
3.7: Update $D'_L = D'_L \cup \{(s_i, y_i)\}_{\beta_i}$, $R' = R' \cup \{r^j : (v^j(\cdot), z^j)\}$, $T = T - T_I - \beta_i \cdot T_R$

Example

Text instance s_i:
"Prime Minister Manmohan Singh today said international environment for India's development was highly favourable..."
Queries:
- Instance label: World
- Rule 1: NGRAM="prime minister" → World ✓
- Rule 2: PROMPT_IS_ABOUT="politics" → World ✓
- Rule 3: NGRAM="international" → World ✗
- Rule 4: –
- Rule 5: –

Table 11: Example from AGNews with $\beta = 5$. All classes are "World," "Sports," "Business," and "Sci/Tech." Out of the rules that were queried, 2 were accepted and 1 was rejected.

INTERVAL outperforms Weak Supervision and Active Learning using fewer expert queries

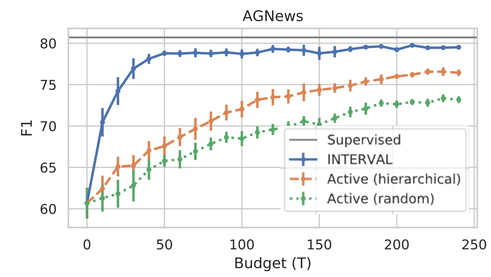
Experimental Results

INTERVAL is more effective than existing Weakly Supervised Learning (WSL) and Active learning approaches even when starting with no expert-written rules.

Method	$\lvert D_L \rvert$	D_U	R	T (T_I, T_R)	YouTube	SMS	IMDB	Yelp	TREC	AGNews	AVG F1
Fully Supervised	100%	–	–	–	94.0	95.6	79.6	87.5	90.3	80.7	88.0
Low Supervised	$20{\cdot}K$	–	–	–	79.8	82.5	61.6	70.4	55.0	58.8	68.0
Semi Supervised	$20{\cdot}K$	✓	–	–	80.7	83.2	63.4	72.0	55.0	60.7	69.2
WSL (ASTRA)	$20{\cdot}K$	✓	✓	–	**90.0**	86.8	**71.2**	80.2	57.0	**75.9**	**76.8**
Active Learning (hierarchical)	$20{\cdot}K$	✓	–	100 (100, 0)	85.3	**89.9**	67.6	**81.2**	**61.4**	71.4	76.1
INTERVAL	$20{\cdot}K$	✓	–	100 (50, 50)	**91.4**	**94.8**	**79.3**	**86.2**	**66.6**	**78.8**	**82.8**

Efficient use of expert feedback

INTERVAL requires as few as $T = 1$ queries to reach $F1$ values that Active Learning cannot match even with $T = 100$ queries.



Rule vs. instance trade-off

Feedback on both rules and instances is more effective than feedback on instances only even when labeling rules are up to 9x more expensive than labeling instances.

Discussion & Future Work

- INTERVAL prompts pre-trained models during training only, and can work with any model for inference, thus **enabling applications where deploying LLMs is impossible**.
- **A more accurate Teacher does not necessarily lead to a more accurate Student.** Would this apply to other Teacher types or distillation types?
- Integrating **richer feedback** (e.g., editing rules) could lead to stronger performance gains.
- Could **LLMs replace experts** without sacrificing accuracy during interactive learning?

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant No. IIS-15-63785.