Cost-efficient Crowdsourcing for Span-based Sequence Labeling: Worker Selection and Data Augmentation

Yujie Wang[1,2*] Chao Huang[3*] Liner Yang[2,4†] Zhixuan Fang[5]
Yaping Huang[1] Yang Liu[2,4] Jingsi Yu[2,4] Erhong Yang[2,4]

[1]Beijing Jiaotong University [2]National Language Resources Monitoring and Research Center for Print Media
[3]The University of California, Davis [4]Beijing Language and Culture University [5]Tsinghua University

Motivation and Task Formulation

Modern deep learning systems rely on **large amounts** of **high-quality** annotated data. **Reliable** data annotation requires costly **expert labor**.

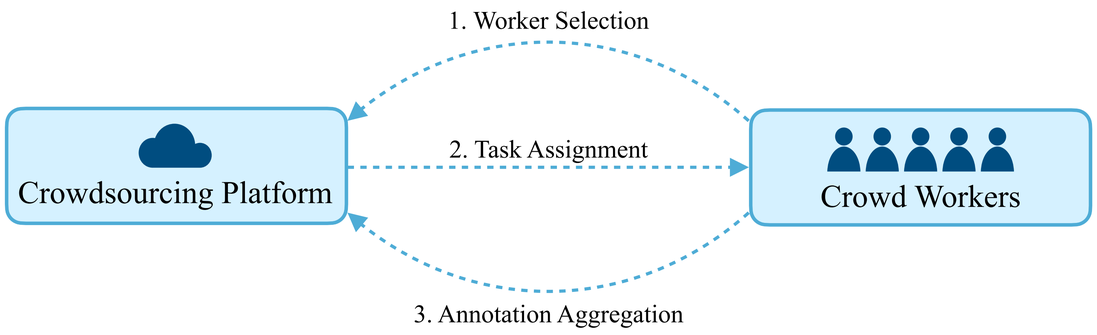
Various Annotation Tasks

Problem 1: How to reduce annotation costs?

- **Crowdsourcing**: Collect a large number of annotations from low-cost crowd workers with less expertise.

Problem 2: How to ensure annotation quality?

- Crowd worker skill levels vary → **Worker selection**.
- Filtering and **aggregating annotations**.



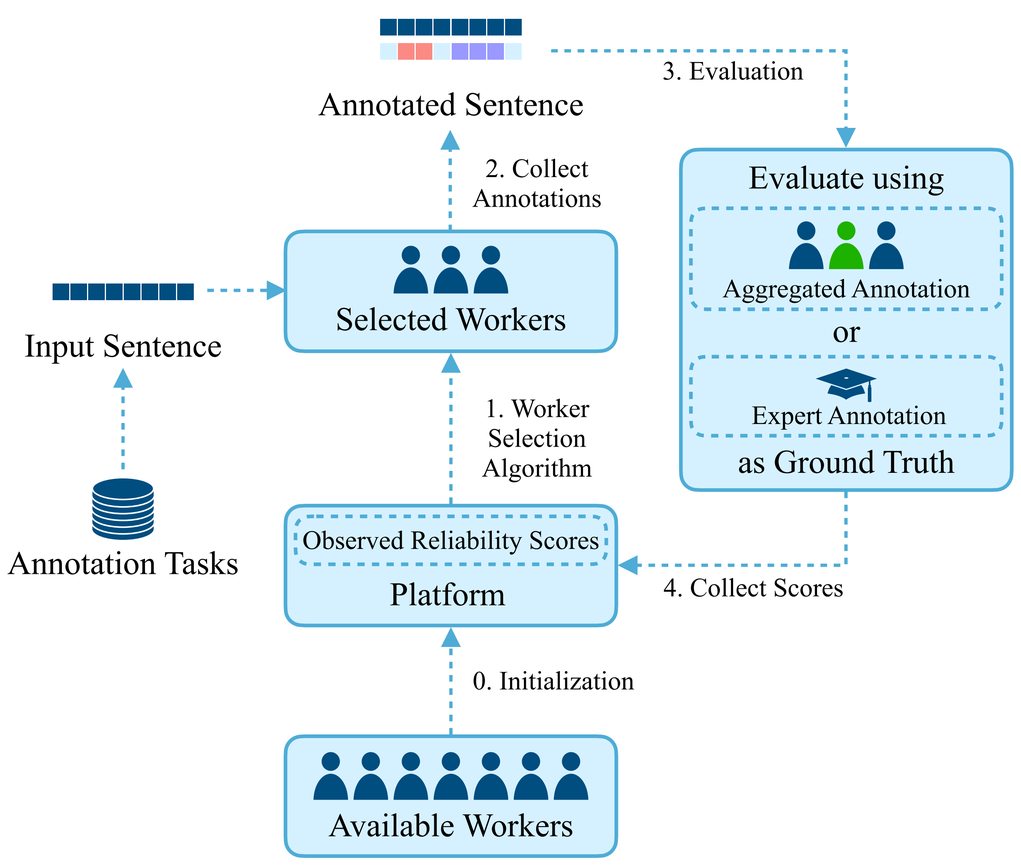
Crowdsourcing Process

Overarching Goal: Obtain high-quality annotations at a lower cost.

Our Online Worker Selection Framework

Problem: How to **select workers** with **better performance** during the crowdsourcing annotation process **dynamically**?

Solution: Proposed **an online worker selection framework** based on the **combinatorial multi-armed bandit** setting**.**



- Select a **subset** of workers in each iteration;
- Use observed worker **reliability scores** as the criterion;
- Based on the **agreement** of the annotations, calculate F_1 reliability scores from **expert feedback** or **aggregated crowd feedback**.

Objective: Maximize the quality (F_1) of the collected annotations.

Data Augmentation Algorithm

Problem: How to do **online** experiments on existing **offline** datasets which does not contain all required annotations?

Solution: Designed a **data augmentation** algorithm via **Shift, Expand and Shrink (SES)** operations on expert annotations.

	Shifting	Expanding	Shrinking
Expert	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm
Modified by 1 word	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm
Modified by 2 words	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm
Modified by 3 words	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm	今天的阳光是轻柔而温暖的 Today's sunshine is gentle and warm

Annotation Generation: Combine and permute all **modified spans** within each sentence to obtain an annotation for the sentence.

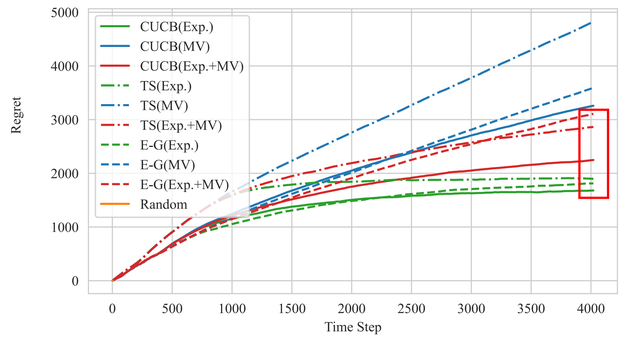
Quality Balancing: Select generated annotations on the sentences for each worker considering their **factual average F1 score**.

Experimental Results

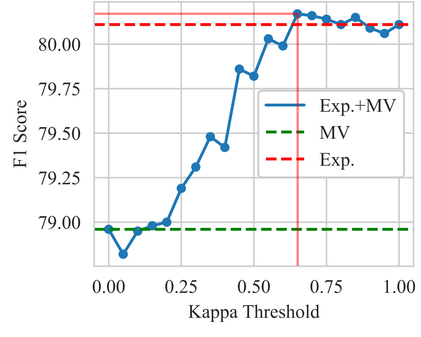
Metric 1 - Regret: Measures the **gap** between the subset of **worker selected** by an algorithm and the **best subset** in each iteration.

Metric 2 - F-score: Measures the **quality** of the collected annotations.

Metric 3 - Cost: Measures the **economic cost** incurred in the process.



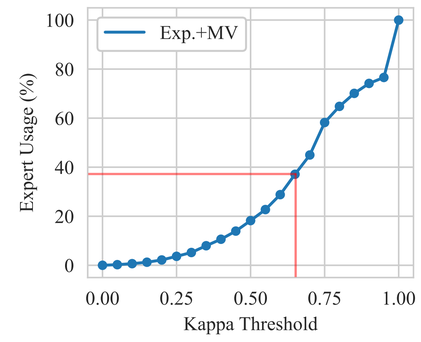
Cumulative Regret



Effect of Kappa on F_1 Score

Method	Token-level			Span-level Exact			Span-level Prop.		
	P	R	F_1	P	R	F_1	P	R	F_1
Oracle	65.69	83.99	70.00	78.15	72.23	74.96	87.97	80.03	83.82
Random	55.95	66.42	57.50	64.42	55.64	59.40	75.70	62.61	68.54
ϵ-G (Exp.)	64.94	80.48	**68.56**	75.24	68.16	**71.34**	85.85	76.79	**81.06**
ϵ-G (MV)	64.44	80.22	67.98	74.69	67.59	70.77	85.67	76.09	80.59
ϵ-G (Exp.+MV)	64.68	80.94	**68.41**	75.08	68.37	**71.40**	85.93	76.62	**81.01**
TS (Exp.)	64.94	79.88	**68.51**	75.64	68.31	**71.57**	85.02	75.71	**80.09**
TS (MV)	64.47	79.19	67.91	74.97	67.54	70.80	84.14	74.21	78.86
TS (Exp.+MV)	64.20	79.09	**67.62**	75.27	67.83	**71.12**	84.77	75.39	**79.81**
CUCB (Exp.)	65.65	80.34	**69.24**	75.94	69.12	**72.20**	86.17	77.22	**81.45**
CUCB (MV)	65.39	80.00	68.91	75.95	68.90	72.08	86.13	76.67	81.12
CUCB (Exp.+MV)	65.33	81.12	**69.11**	75.70	69.30	**72.21**	86.17	77.28	**81.48**

P, R, F_1 Scores



Effect of Kappa on Cost

Observation: The proposed method matches expert annotation quality while significantly reducing costs.

Conclusion

- It presents the exploration of **worker selection for span-based sequence labeling** tasks, recognizing the unique challenges.
- It employs the span-level F1 score, **evaluated by experts and crowd workers**, as a feedback mechanism, for accurate worker selection.
- It introduces a **data augmentation technique** to counteract the limitations of real datasets, enabling effective offline simulations.
- Rigorous experiments demonstrate the **efficacy** of the proposed method, achieving **impressive F1 scores** while **significantly reducing expert annotation costs**.

GitHub: https://github.com/**blcuicall/nlp-crowdsourcing**