Exploring the Impact of Sub-Task Inter-Dependency on Crowdsourced Event Annotation

Tianyi Li, [1] **Ping Wang,** [2] **Tian Shi,**[3] **Andrey Esakia** [4]

[1] Purdue University
[2] Stevens Institute of Technology
[3] Moody's Analytics
[4] Virginia Tech
li4251@purdue.edu, ping.wang@stevens.edu, researchtianshi@gmail.com, esakia@cs.vt.edu

Abstract

Unlike most homogeneous labeling tasks, annotating Event Extraction (EE) datasets involves four inter-dependent sub-tasks. Making a mistake in one sub-task may impact the accuracy of other sub-tasks. As a result, it is challenging and error-prone to crowdsource event annotation. In this work, we explore how sub-task inter-dependency may impact and facilitate crowdsourced event extraction annotation.

Introduction

Event extraction (EE) is a natural language processing (NLP) problem that aims to detect and retrieve attributes of real-world events from unstructured natural language texts. For a sentence where an event is mentioned, there needs to be a *trigger word* that indicates the occurrence of an event, a corresponding *event type*, the *arguments* involved in this event, and the *roles* of each event argument. Table 1 shows a sentence describing an event where a person was born.

Jane Doe was born in Casper, Wyoming on March 18, 1964.

T1	"born"		
T2	BE-BORN		
T3	"Jane Doe"	"Casper, Wyoming"	"March 18, 1964"
T4	Person-Arg	Place-Arg	Time-Arg

Table 1: An example of event annotation. For an unstructured sentence, four annotation tasks are conducted. **T1** identifies the trigger word ("born"). **T2** classifies the event type (BE-BORN) of the trigger word identified in T1. **T3** identifies the arguments of the event classified in T2. **T4** classifies the role of each argument identified in T3.

Extracting event attributes from natural language data is fundamentally challenging due to the abundance of vague and indefinite expressions (Ludlow 1999). Taking closed-domain event extraction as an example, there are *event schemata* that define the rules for "taggability" (Walker et al. 2006), that is, what event types and event attributes should be considered for annotation. Successful annotation requires one to grasp the event schemata, make sense of the sentences, and annotate the event information accordingly.

Crowdsourcing has been used to annotate data for various NLP problems, such as named entity recognition (Wang et al. 2012), topic categorization (Chilton et al. 2013), and sentiment classification (Brew, Greene, and Cunningham 2010). However, the application of crowdsourcing in event annotation is still in its early stage. The use of non-expert crowds has been explored in the context of *verifying* existing entity relation annotations (Liu et al. 2016; Callison-Burch 2009), but it remains unclear whether and how they can *create* a series of inter-dependent annotations. Some newly developed data sets used crowdsourcing to annotate trigger words (T1) and event types (T2) (Wang et al. 2020), but there were insufficient details about the crowdsourcing methods and the resulting annotations were found to have mixed-quality (Zhang et al. 2022).

To reduce the burden of expert annotators, and support larger scale event annotations, we explore the use of paid, micro-task crowdsourcing to conduct all four sub-tasks of event annotation. Prior work shows that additional context can improve crowd performance in sensmaking tasks but may incur cognitive overload that diminishes the analysis quality (Li et al. 2019; Alagarai Sampath, Rajeshuni, and Indurkhya 2014). We draw inspiration from the prior work and investigate the potential of using related annotation sub-tasks as meaningful contexts to onboard and scaffold novice crowds in event annotation. We also explore how the additional sub-tasks may impact the annotation quality and the crowds' perception of workload.

Approach

We conducted a between-subject experiment on Amazon Mechanical Turk to compare the impact of sub-task inter-dependency on the event annotation quality and workload.

Experiment Design We considered three event annotation workflows, each with a different level of sub-task combination (Table 2): **L1.** each crowd worker only works on one event annotation sub-task; **L2.** each crowd worker works on two event annotation sub-tasks: T1 and T2 (event detection), or T3 and T4 (event argument detection); **L3.** each crowd worker works on all four sub-tasks together.

To mitigate the confounding factors such as individual differences among crowd workers and sentences, each crowd task contains 10 different sentences to annotate and

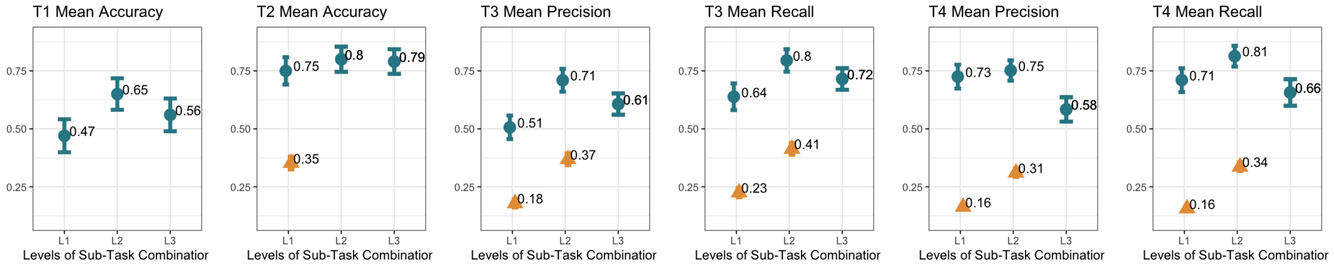

Figure 1: Task Performance of each annotation task. The blue round dots represent the mean task performance scores. We also show the adjusted performance to account for error propagation with yellow triangle dots.

was assigned to 10 different crowd workers. The 10 sentences were selected from the ACE event annotation guidelines (Linguistic Data Consortium 2005). The ACE dataset is one of the most well-validated event annotation datasets (Li et al. 2021) and the ACE guideline (Linguistic Data Consortium 2005) contains detailed definitions and rules for each event type, as well as examples with explanations. We use the ACE annotations as the ground truth to assess the quality of the crowdsourced event annotations.

Levels of Sub-Task Combination	Crowd Tasks			
One Sub-Task per Worker	T1	T2	T3	T4
Two Sub-Tasks per Worker	T1+T2		T3+T4	
All Sub-Tasks per Worker	T1+T2+T3+T4			

Table 2: The three levels of sub-task combination and the crowd tasks in each level.

Participants We hired novice crowd workers from Amazon Mechanical Turk (MTurk) with the criteria of completing more than 100 Human Intelligence Tasks (HITs) with above 95% acceptance rate. We estimate the time needed to complete each crowd task with a pilot study. Based on our local minimum wage, each HIT pays $1.2. We also incentivize crowd workers with a $1 bonus when more than 8 out of the 10 sentences are annotated correctly.

Procedure Each crowd task contains four phases: 1) introduction and examples, 2) training tasks, 3) annotation tasks (10 sentences), and 4) post-task questionnaire. After accepting the HIT, the crowd workers can withdraw by returning the HIT at any time. Otherwise, each crowd worker will be guided through the four phases.

Results and Discussions

We recruited 70 crowd workers from MTurk (10 for each crowd task), 29 were female (41.4%), 36 were male (51.4%), and five preferred not to tell (7.2%). Most of the crowd workers were between the age 25-65 (N = 63, 90%), two participants were between the age 18-24, and five participants preferred not to reveal their age.

Impact on Annotation Quality The annotation quality results are shown in Figure 1. Since each event mention has one trigger word (T1) and one event type (T2), and could have multiple event arguments (T3) with different argument roles (T4), we use *accuracy* to measured the annotation quality of T1 and T2, and *precision* and *recall* to measure the annotation quality of T3 and T4.

We analyzed the annotation performance with a mixed-effect model (Bates et al. 2014), where the three *levels* of sub-task combination serve as the main effect. We also examined if different *event types* (blocking factor) and individual *sentences* (random effect) had significant impact on the crowd performance. The analysis results show that task context levels (main effect) significantly influenced the performance of T1, T3, and T4. Working on inter-dependent sub-tasks has led to higher annotation quality than when each sub-task is annotated independently. The event type (blocking effect) and individual sentences (random effect) did not significantly influence the performance of any annotation tasks. In the post-task questionnaire, 50% of the crowd workers rated their success in accomplishing the annotations (*Performance*) positively (with ratings above neutral).

Impact on Annotation Workload We measure the crowd task workload with two metrics: the self-reported perception using the NASA-TLX survey (Hart and Staveland 1988) and the HIT elapsed time on MTurk.

Overall, the crowd perceives the crowd tasks as requiring hard work (*Effort*) and mentally demanding (*Mental Demand*), regardless of the crowd task. Surprisingly, the Chi-squared Test of Independence suggests that crowd perceptions are independent of the task conditions. Thus, the crowds did not perceive the crowd tasks differently, despite that the crowd tasks had different types and different numbers of annotation sub-tasks. In fact, we saw a low agreement of workload perception among different crowd workers (Krippendorff's alpha less than 0.1). In other words, the perception of workload varies more among individual crowd workers than across our controlled conditions.

Discussion and Future Work Our initial results show that novice crowds perform better on event annotation tasks when working on multiple inter-dependent sub-tasks together, without perceiving additional workload. We plan to verify this insight with additional data collection and more in-depth data analysis. Figure 1 shows an adjusted performance score (the orange triangle dots) by penalizing the performance scores of T2, T3, and T4 with the observed annotation quality in the preceding tasks.

References

Alagarai Sampath, H.; Rajeshuni, R.; and Indurkhya, B. 2014. Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 3665–3674. New York, NY, USA: Association for Computing Machinery. ISBN 9781450324731.

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Brew, A.; Greene, D.; and Cunningham, P. 2010. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, 145–150. NLD: IOS Press. ISBN 9781607506058.

Callison-Burch, C. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286–295. Singapore: Association for Computational Linguistics.

Chilton, L. B.; Little, G.; Edge, D.; Weld, D. S.; and Landay, J. A. 2013. Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 1999–2008. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318990.

Hart, S. G.; and Staveland, L. E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, P. A.; and Meshkati, N., eds., *Human Mental Workload*, volume 52 of *Advances in Psychology*, 139–183. North-Holland.

Li, Q.; Li, J.; Sheng, J.; Cui, S.; Wu, J.; Hei, Y.; Peng, H.; Guo, S.; Wang, L.; Beheshti, A.; and Yu, P. S. 2021. A Compact Survey on Event Extraction: Approaches and Applications.

Li, T.; Manns, C. J.; North, C.; and Luther, K. 2019. Dropping the Baton? Understanding Errors and Bottlenecks in a Crowdsourced Sensemaking Pipeline. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Linguistic Data Consortium. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*, 5.4.3 2005.07.01 edition.

Liu, A.; Soderland, S.; Bragg, J.; Lin, C. H.; Ling, X.; and Weld, D. S. 2016. Effective Crowd Annotation for Relation Extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 897–906. San Diego, California: Association for Computational Linguistics.

Ludlow, P. 1999. *Semantics, tense, and time: an essay in the metaphysics of natural language*. MIT Press.

Walker, C.; Strassel, S.; Medero, J.; and Maeda, K. 2006. *ACE 2005 Multilingual Training Corpus*.

Wang, J.; Kraska, T.; Franklin, M. J.; and Feng, J. 2012. CrowdER: Crowdsourcing Entity Resolution. *Proc. VLDB Endow.*, 5(11): 1483–1494.

Wang, X.; Wang, Z.; Han, X.; Jiang, W.; Han, R.; Liu, Z.; Li, J.; Li, P.; Lin, Y.; and Zhou, J. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1652–1671. Online: Association for Computational Linguistics.

Zhang, W.; Ingale, B.; Shabir, H.; Li, T.; Shi, T.; and Wang, P. 2022. Event Detection Explorer: An Interactive Tool for Event Detection Exploration. *arXiv preprint arXiv:2204.12456*.