Disentangling Inherent Ambiguity and Disagreement in Crowdsourced Annotation

Quanze Chen
University of Washington
Paul G. Allen School of Computer Science and Engineering
Seattle, WA
cqz@cs.washington.edu

Abstract

Crowdsourced annotations have become a crucial resource for training and evaluating machine learning systems. However, with the increasing complexity of tasks comes a growing demand for annotations that involve aspects of data inherently ambiguous or where humans are uncertain of how to make judgments. Traditional elicitation methods focuses on the idea of collecting single precise judgments from each annotator, but this can misrepresent the sources of uncertainty—an individual annotator may be uncertain about an item or different annotators may disagree. My research will aim to addresses these problems through two directions: (1) Can we design novel elicitation mechanisms that can disentangle ambiguity from disagreement on subjective annotation tasks where both forms of uncertainty manifest; and (2) Can we design feedback loops in decision processes that are aware of human uncertainty and provide targeted solutions to address it.

Crowdsourcing has been used for a wide variety of tasks ranging from labelling images to judging toxicity of text. However, as soon as sufficient annotations are collected, it becomes apparent that not all annotators will label the same item in the same way. In fact, on subjective tasks like evaluating toxicity, the same annotator may not even agree with their own past annotations of the same item (Gordon et al. 2021b). A long running challenge in crowdsourced annotation of data has been to understand various factors that contribute to uncertainty in crowdsourced data and reducing the impact of uncertainty (Aroyo et al. 2019).

One source of uncertainty reflects through disagreement between various crowd workers. Disagreement can arise from simple reasons such as random mistakes or fatigue to complex sources like disagreement on scales or fundamental differences in mental models. For example, when judging toxicity of online comments, workers' diverse backgrounds may cause them to disagree about whether comments are toxic or not and what different levels of toxicity entail based on their description. In addition to disagreement, uncertainty and inconsistency can also ariss from ambiguity inherent to the items being annotated. While a worker may be able to tell an incoherent translation from a perfect one, they might have trouble deciding whether it's worse to have bad grammar or missing content in a translation. With these varied sources of uncertainty in mind, we can see that the traditional methods to elicit huamn judgemtsn—like semantic (Likert) rating—can often worsen uncertainty, especially if they assume a single answer is always possible.

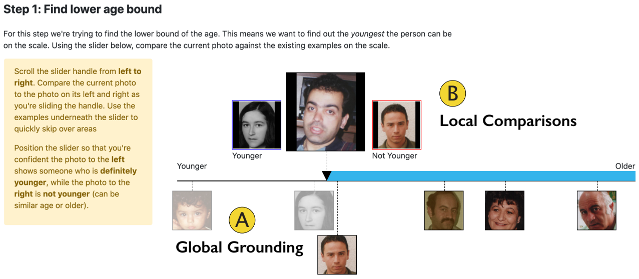
Related Work

There is a growing demand for human annotation in domains involving ambiguous or subjective examples, largely due to rapid progress in machine learning. Human rating annotation has been used to create or validate a variety of training data. For example, in the domains investigating toxicity (Wulczyn, Thain, and Dixon 2017), misinformation and credibility (Bhuiyan et al. 2020; Mitra and Gilbert 2015), and emotionally manipulative text (Huffaker et al. 2020). At the same time, there is also increasing concern for the robustness of datasets collected (Welty, Paritosh, and Aroyo 2019) and whether nuances like uncertainty are being represented (Aroyo and Welty 2015).

In the past, many have looked at this problem from a quality control perspective. Aggregation approaches like majority vote (Snow et al. 2008) or EM (Dawid and Skene 1979; Whitehill et al. 2009; Welinder et al. 2010) were proposed early on to account for disagreement as noise. However, as tasks shifted to more challenging domains, noise became a less important factor in disagreement compared to genuine divergences in task interpretation. In objective domains, rubrics (Yuan et al. 2016) and training (such as via gated instructions (Liu et al. 2016)) have been proposed as effective ways to unify understanding, though these methods can require significant time investment in development. One other prior line of work, structured labeling (Kulesza et al. 2014; Chang, Amershi, and Kamar 2017), proposes tools and techniques designed to assist the development of shared understanding from the ground up through collaboratively creating and refining taxonomies.

As a more flexible way to address disagreement, deliberation (Drapeau et al. 2016; Schaekermann et al. 2018; Chen et al. 2019) through argumentation has also been proposed in various forms to directly resolve disagreement. However, deliberation methods can be costly and when uncertainty arises from ambiguity rather than disagreement, deliberation may fail to provide benefits.

More recent lines of work are also recognizing the defi-



Step 1: Find lower age bound

For this step we're trying to find the lower bound of the age. This means we want to find out the *youngest* the person can be on the scale. Using the slider below, compare the current photo against the existing examples on the scale.

Scroll the slider handle from **left to right**. Compare the current photo to the photo on its left and right as you're sliding the handle. Use the examples underneath the slider to quickly skip over areas

Position the slider so that you're confident the photo to the **left** shows someone who is **definitely younger**, while the photo to the **right** is **not younger** (can be similar age or older).

Figure 1: Diagram illustrating the main features of Goldilocks' range-based annotation process.

ciencies in labels that assume items can be judged with certainty, proposing instead to use answer distributions in the form of allowing multiple labels (Jurgens 2013; Dumitrache 2015; Dumitrache, Aroyo, and Welty 2018) to capture the sources of uncertainty rather than attempt to remove it. Metrics based on inter-annotator agreement have also been proposed as a way to check the quality and consistency of an annotation task's results (Welty, Paritosh, and Aroyo 2019) which may inform downstream use of the data. Uncertainty in human annotations has also started to be recognized in model performance evaluation (Gordon et al. 2021a) and other downstream tasks that make use of increasinly uncertain human judgments.

Research Questions

My research focuses on answering the following research questions:

- **RQ1**: How can we achieve better consistency on subjective annotation tasks?
- **RQ2**: How can we capture varied source of uncertainty in an intuitive way?
- **RQ3**: How can we design human-in-the-loop workflows that make use of disentangled ambiguity and disagreement to improve consensus and represent diverse perspectives?

Proposed Experiments and Progress So Far

In previous work, we have made some progress exploring RQ1 and 2 by building a new workflow to improve upon traditional scalar rating to simultaneously capture disagreement and ambiguity which we have published in CSCW '21 (Chen, Weld, and Zhang 2021). In the follow up work, we plan to explore RQ3, looking at how using multi-dimensional of uncertainty can be used in decision processes to improve consensus. We will use data collected through Goldilocks to evaluate targeted methods to reduce uncertainty. We will also explore uncertainty-aware ensemble models that can simulate diverse sub-population disagreements rather than providing only a global prediction that may hide diverse and alternate interpretations.

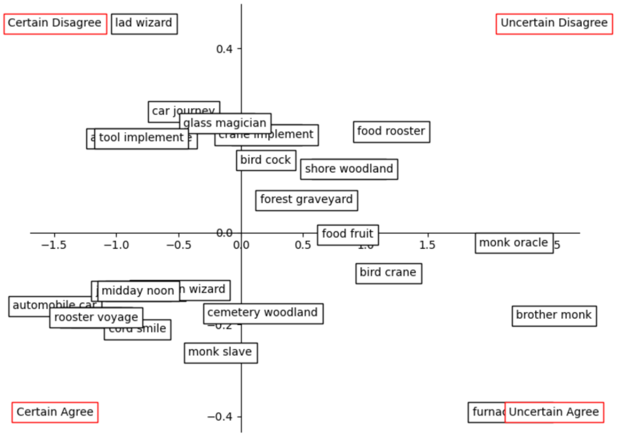


Figure 2: Preliminary results illustrating the separation of ambiguity (X-axis) and disagreement (Y-axis) relative to the average on a word similarity task (WordSim353).

Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty

Traditional annotation methods lump sources of uncertainty into a single measurement like inter-annotator agreement. However this offers limited insight into why a group of annotators may be uncertain—is it because the item itself has multiple ambiguous interpretations, or because the annotators disagree with each other on what the correct evaluation should be. In Goldilocks we set out to directly elicit where the uncertainty is coming from by creating a novel interface that uses a two-step bounding process to elicit an evaluation of inherent ambiguity from each annotator in the form of a range (Figure 1). To facilitate this, we also incorporated past annotations as anchor examples to ground otherwise vague scales and enable comparisons on otherwise absolute scales. Annotation experiments conducted with Goldilocks showed that using our approach does improve consistency consistency (RQ1) on the more subjective task domains (comment toxicity, food satiety) and that the ranges collected we collect as a resulte better modelled the sources of uncertainty producing pairwise comparison distributions that more closely aligned to ground truth.

Next Steps: Uncertainty Aware Decision Processes and Models

Following this work, we want to explore whether we can address and reduce the uncertainty once we are aware of its components through an uncertainty-aware decision process. Figure 2 shows an example where ambiguity and disagreement were separated on a word similarity task similar to what can be collected through Goldilocks. Here high ambiguity examples seem to align with words that have multiple word senses while high disagreement examples align with different understanding about what it means to be similar. By looking at where a word pair falls in the space, we might choose to apply different interventions—providing additional context such as using the words in a sentence when the word senses are ambiguous, or using deliberation to refine how workers reason about establishing whether words are similar.

To explore this idea, we plan to conduct simulation studies based on several tasks involving subjectivity. We will collect disentangled uncertainty measurements in the form of Goldilocks ranges for each item. We plan to then individually apply separate methods to reduce uncertainty: deliberation, adding context and changing rubrics. If the source of uncertainty matches the intervention, uncertainty should be reduced significantly for the instance (reflected through data points moving in the uncertainty space in respons to interventions). We will also simulate various decision processes by buidling hypothetical datasets if they were constructed through the process to evaluate aspects like efficiency and cost.

Finally we plan to explore the feasibility of ensemble models trained over groups of ranges as a possibility to predict multi-source uncertainty and automate some of these decision processes on unseen examples. Such a hypothetical model may also offer improved fairness through the ability to provide dissenting opinions by itself.

References

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, 1100–1105.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36(1): 15–24.

Bhuiyan, M. M.; Zhang, A. X.; Sehat, C. M.; and Mitra, T. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proceedings of the ACM on Human-Computer Interaction* (CSCW).

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2334–2346.

Chen, Q.; Bragg, J.; Chilton, L. B.; and Weld, D. S. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Chen, Q.; Weld, D.; and Zhang, A. X. 2021. Goldilocks: Consistent Crowdsourced Scalar Annotations with Relative Uncertainty. In *ACM CSCW*.

Dawid, A.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm. *Applied Statistics* 28(1): 20–28.

Drapeau, R.; Chilton, L. B.; Bragg, J.; and Weld, D. S. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Dumitrache, A. 2015. Crowdsourcing Disagreement for Collecting Semantic Annotation. In *ESWC*.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. In *HCOMP*.

Gordon, M.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. 2021a. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Gordon, M. L.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. S. 2021b. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* .

Huffaker, J. S.; Kummerfeld, J. K.; Lasecki, W. S.; and Ackerman, M. 2020. Crowdsourced Detection of Emotionally Manipulative Language. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* .

Jurgens, D. 2013. Embracing Ambiguity: A Comparison of Annotation Methodologies for Crowdsourcing Word Sense Labels. In *HLT-NAACL*.

Kulesza, T.; Amershi, S.; Caruana, R.; Fisher, D.; and Charles, D. 2014. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3075–3084.

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 NAACL and HLT 2016*.

Mitra, T.; and Gilbert, E. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

Schaekermann, M.; Goh, J.; Larson, K.; and Law, E. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1–19.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *2008 Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. J. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems (NIPS)*, 2424–2432.

Welty, C.; Paritosh, P.; and Aroyo, L. 2019. Metrology for AI: From Benchmarks to Instruments. *arXiv preprint arXiv:1911.01875* .

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Laberlers of Unknown Expertise. In *In Proc. of NIPS*, 2035–2043.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Yuan, A.; Luther, K.; Krause, M.; Vennix, S. I.; Dow, S. P.; and Hartmann, B. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, 1005–1017. New York, NY, USA: Association for Computing Machinery.