Error-Correction and Aggregation in Crowd-Sourcing of Geopolitical Incident Information

Alexander G. Ororbia II[1](✉), Yang Xu[1], Vito D'Orazio[2], and David Reitter[1]

[1] Pennsylvania State University, University Park, USA
ago109@ist.psu.edu
[2] Harvard University, Cambridge, USA

Abstract. A discriminative model is presented for crowd-sourcing the annotation of news stories to produce a structured dataset about incidents involving militarized disputes between nation-states. We used a question tree to gather partially redundant data from each crowd worker. A lattice of Bayesian Networks was then applied to error correct the individual worker annotations, the results of which were then aggregated via majority voting. The resulting hybrid model outperformed comparable, state-of-the-art aggregation models in both accuracy and computational scalability.

1 Introduction

Crowd-sourcing has challenged the notion that complicated problems call for great expertise. Instead, it parallelizes the solution-finding process: many untrained individuals contribute to a joint solution. Tasks ranging from natural language processing (NLP) [14] to image recognition [15] have been shown to be amenable to the use of crowds in lieu of experts. Given these successes, the use of crowd-sourcing has proliferated to other fields of study, notably the quantitative social sciences [1,7].

In social science research, where metrics for studying social phenomena are often derived by expert judgment and analysis, crowd-sourcing has the potential for ubiquitous application. For example, *militarized conflict* has traditionally been measured by the expert analysis of text documents [10]. In sufficient numbers, however, non-experts should be able to analyze these documents as effectively as experts. This leaves a problem of aggregation: how can redundant work be most effectively combined?

In this paper, we evaluate methods for aggregating partially redundant information from crowd workers to code geopolitical incidents using the criteria defined by the Militarized Interstate Dispute (MID) project [10]. To begin, we deconstructed the task into several simple and objective questions, the answers to which provided us with sufficient information to annotate the document. Unlike previous approaches, we asked partially redundant questions that do not follow a one-to-one mapping to target variables. The data gathered from workers inform

N. Agarwal et al. (Eds.): SBP 2015, LNCS 9021, pp. 381–387, 2015.
DOI: 10.1007/978-3-319-16268-3_47

a set of Bayesian Networks, which are trained to error-correct individual worker results. The models are combined with a voting scheme that aggregates multiple worker inputs to perform different classification tasks. Finally, we compare the accuracy of our approach with competing models and find that our hybrid approach outperforms all others.

2 Related Work

The coupling of human and machine intelligence is emerging as a critical tool in utilizing large-scale datasets where manual labeling is expensive [3]. Often, machine learning algorithms are trained to emulate collective human intelligence through the interaction with human users [8]. Both applications [2] and evaluations of efficiency and cost effectiveness [11] are available. Hybrid methods have proven effective in handling issues of global interest, such as early stage tracking of disease outbreaks [9]. The successes reported in these studies motivate the current study.

Benchmark platforms such as SQUARE [13] have made available representative worker aggregation methods that improve upon simple majority voting. We compete with two aggregation models, the state-of-the-art ZenCrowd [5] and the established Dawid and Skene & Naive Bayes method [4] (DS/NB). [1] Both ZenCrowd and DS/NB model worker behavior and problem difficulty to vertically aggregate responses. In contrast, our method circumvents the computational overhead imposed by models of behavior and task complexity. While task-agnostic, it can still incorporate domain heuristics.

3 Methodology

3.1 Crowd-Sourcing

Workers on Amazon Mechanical Turk (AMT) read a news story and answered a set of simple, objective questions about it implemented using Qualtrics. Questions were designed to address the coding criteria defined by the MID project, a well-known, ongoing effort that collects data on international conflict [10]. News stories were randomly sampled from a set of "potentially relevant" MID documents in equal portions from years 2007, 2009, and 2010.[2] Each news story read by the workers was either irrelevant or about a *threat*, *display*, or *use* of military force. There are three primary coding tasks, each specifying a target variable: 1) the hostility level (threat/display/use of force), 2) the initiator and target nation-states, and 3) the type of actions taken by these countries.[3]

[1] A third, and similar to our own work, is [12], where the DALE model was proposed to solve the object localization task. No public implementation was available.

[2] The algorithm in [6] was used to create the set of "potentially relevant" source documents, from which 150 were randomly sampled per year. After discarding some for formatting reasons, 446 total were left.

[3] See [10] for additional details about the MID coding ontology.

Table 1. Sample features provided by workers

Meaning	Value Examples
Hostility level of MII	Non-incident, Threat to use force, Display of force, Use of force
Type of action taken by country that started incident	Alert, Seizure, Attack, Join interstate war etc.
Is action just verbal, or material?	Verbal, Material
Whether or not a story is conflictual or cooperative	Cooperative, Conflictual
State entity first taking action (the ”initiator”)	Afghanistan, Armenia, etc.
State entity opposing the initiator	Afghanistan, Armenia, etc.

3.2 The Question Tree

A question tree, comprised of blocks of multiple-choice and yes-or-no questions, was used to guide the workers to finish the coding tasks.[4] Workers are branched to different sub-blocks depending on their answers to previous questions. The answers to many of these questions provide information necessary for completing the underlying coding tasks and building feature representations for horizontal integration. Additional questions extract partially redundant information that is later used for error correction. Examples of the features provided by the questionnaire are shown in Table 1.

1644 workers on Mechanical Turk coded the 446 documents. While 1251 workers did this task only once, some completed many dozens of annotation tasks. On average, each document was coded by 8.47 different workers (range: 6–10).

3.3 Prediction Targets

There are 5 labels that are most informative of the nature of an MII and hence are the target variables that we predict. *Initiator* refers to the country that took the first action in the dispute. *Type* is the primary action type taken by the initiator. *Target* refers to the country that is the target of the initiator's action. *Level* refers to the hostility level of the action taken by the initiator. *Incident* is a binary variable that distinguishes a story about a militarized conflict event between nation-states from other articles. All target variables but *Incident* are directly answered by corresponding regular questions given to the workers. *Incident* is built from answers to these questions. The error-correction approach will, however, use all available information to revise those choices.

For our gold standard to compare each model's predictions against, each document was independently labeled by three subject matter experts, each of whom were graduate students of political science experienced with the MID coding scheme. Disagreements among the experts were resolved by majority

[4] A copy of the Qualtrics questionnaire is available at http://goo.gl/TZnkVd.

vote. In 19 cases, all three disagreed and these were resolved by subsequent discussions of MID coding rules.[5]

3.4 Approaches to Predicting Targets

Several approaches for predicting the target classes were evaluated. The most frequent class label served as a baseline (*Baseline*) while the first model, *Voting*, performs a vertical aggregation across worker annotations via their modal response. This commonly-used approach only aggregates direct responses for single variables. The same limitation holds for the two previously proposed approaches we also compare our methods against, ZenCrowd (*Zen*) and DS/NB (*Bayes*).

The *Horizontal* approach consists of a classifier trained on annotated story features as determined by the workers. Information from all features is integrated to predict all variables, separately for each worker.

Finally, our hybrid method *Hori+Vert* first trains a set of classifiers on a training set of worker-annotated story features. The resulting model is applied to each worker's annotations of a single story. This produces a prediction matrix for each unseen story, where columns represent the five predicted variables and rows correspond to error-corrected annotations of the workers. The most-frequent choice is then computed for each column, yielding the predictions for each story. At this stage, simple domain heuristics may be applied to guarantee plausibility. (For example a "non-incident" never has initiators or targets.) Ties are resolved at random.

The Bayesian (Belief) Network, a probabilistic directed graphical model, was chosen as the base classifier for both the *Horizontal* and *Hori+Vert* models [6]. Such a model does not require explicit hyper-parameter tuning and can be constructed efficiently, advantages we exploit in composing our lattice of discriminative experts.

4 Results

Evaluation results were produced on non-overlapping test sets, separate from the training data. We used 40-fold cross-validation by story (i.e., classifiers were never trained on the same story used to evaluate them). As shown in Table 2 and Figure 1, the results of our experiments indicate that a combined approach, which leverages the power of crowd-sourcing aggregation and supervised machine learning integration, yields the best predictive model for each of the

[5] Inter-annotator agreement was acceptable (Cohen's Kappa across years: 0.7-0.78). While this indicates a well-defined coding scheme, it also shows that MIDs remain difficult to code.

[6] Waikato Environment for Knowledge Analysis (3.7.10) was used to build the models. This model performed best in comparison to other algorithms we tuned, such as the Support Vector Machine (linear & Gaussian kernels).

Table 2. Predicting target variables via different approaches

	Baseline	Hori+Vert	Horizontal	Voting	Worker	Bayes	Zen
Initiator	11.38	**73.75**	64.99	69.06	57.18	67.43	64.10
Target	14.49	**71.25**	60.94	68.51	56.28	64.87	63.58
Type	46.47	**73.33**	60.45	68.97	51.82	64.10	67.50
Level	33.05	**83.33**	69.99	71.26	59.81	69.23	68.21
Incident	53.53	**87.50**	77.27	−−	−−	−−	−−

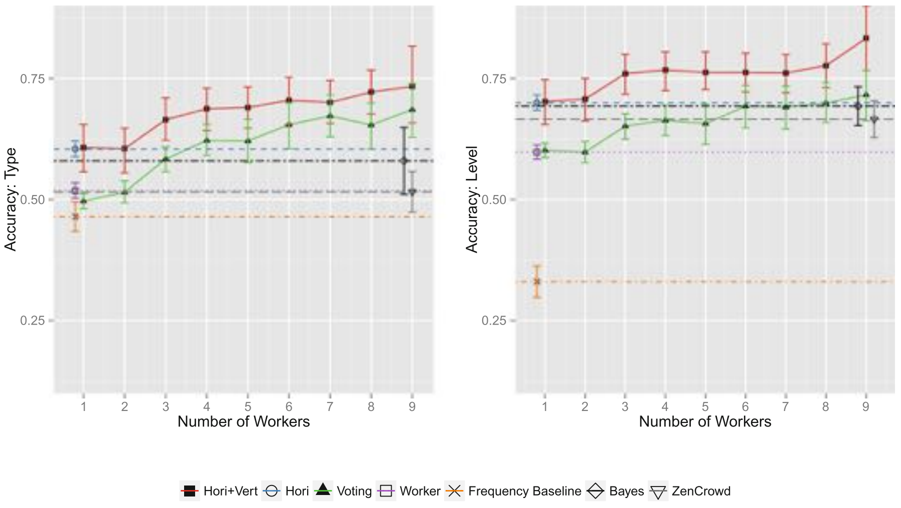

Fig. 1. Accuracy vs. Number of Workers. Some data from workers annotating very few or many stories were removed for this analysis.

target variables.[7] Furthermore, the performance of the *Zen* and *Bayes* aggregation models worsened with task complexity. Specifically, the results indicate that these models do not scale well with the size of the category set (confirming a hypothesis stated in [13]). Run-time performance similarly worsens as problem difficulty increases. For example, when predicting *Initiator*, the average model training time is 18.97 seconds for *Zen* and 6.62 for *Bayes*, as compared to 0.008 for *Hori+Vert*. Similar trends are exhibited for each target variable.[8]

Our hybrid model continues to outperform the non-hybrid ones as additional workers are employed. This is seen in Figure 1 (*Hori+Vert*). Intuitively, this makes sense since our model leverages both the horizontal *and* vertical features of the data, while the other models are restricted to one or the other.

[7] Since workers were not asked to directly classify a story as MID or non-MID, for *Incident* our hybrid model was only compared to the *Horizontal* and *Baseline* models.

[8] For *Target*, average model training time is 19.17 seconds for *Zen*, 4.73 for *Bayes*, and 0.009 for *Hori+Vert*. For *Type*, it is 1.76 seconds for *Zen*, 0.56 for *Bayes*, and 0.005 for *Hori+Vert*. For *Level*, it is 0.14 for *Zen*, 0.12 for *Bayes*, and 0.004 for *Hori+Vert*.

One explanation for why the *Hori+Vert* model outperforms the others pertains to the influence of erroneous annotations. Specifically, some workers will very likely make erroneous annotations, and the modal computation step helps to mitigate the impacts of such mistakes. In basic voting or other aggregation models (Bayes, Zen), however, erroneous annotations are still leveraged for cross-feature prediction.

5 Conclusion

Geopolitical incident news stories were annotated by non-expert workers according to the MID project coding rules. The prediction ability, when using partially redundant information provided by the workers, of various algorithms was then evaluated.

The overall performance of our error-correcting lattice of Bayesian Networks outperforms aggregation or classification-only methods. The advantage of our approach is that it integrates annotations horizontally via supervised learning, and vertically aggregates the results via the predicted majority vote for a group of workers examining a given story. The ensemble nature of our hybrid approach allows for an additional level of error-correction, yielding a model that not only takes into account relationships between features but also target predictor values. In future work, simple rule-mining could be used to "tune" these higher-level correction heuristics to the target task. Our method exploits even workers that make mistakes by integrating across their answers to both direct and indirect questions without the need for modeling worker behavior.

Acknowledgments. We acknowledge seed funding from the Social Sciences Research Institute. We thank Jessie Li, C. Lee Giles, project collaborators Glenn Palmer, Michael Schmierbach, & data annotators Matthew Lane, & Michael Kenwick.

References

1. Benoit, K., Conway, D., Laver, M., Mikhaylov, S.: Crowd-sourced data coding for the social sciences: Massive non-expert human coding of political texts. Presentation at the 3rd Annual New Directions in Analyzing Text as Data Conference. Harvard University (2012)
2. Boia, M., Musat, C.C., Faltings, B.: Acquiring commonsense knowledge for sentiment analysis through human computation. In: 28th American Association for Artificial Intelligence (2014)
3. Culotta, A., McCallum, A.: Reducing labeling effort for structured prediction tasks. In: Proceedings of the 20th National Conference on Artificial Intelligence, AAAI 2005, vol. 2, pp. 746–751. AAAI Press (2005)
4. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 20–28 (1979)
5. Demartini, G., Difallah, D.E., Cudr-Mauroux, P.: ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: Proc. 21st International Conference on World Wide Web, pp. 469–478. ACM (2012)

6. D'Orazio, V., Landis, S.T., Palmer, G., Schrodt, P.: Separating the wheat from the chaff: Applications of automated document classification using support vector machines. Political Analysis **22**(2), 224–242 (2014)
7. Gao, H., Wang, X., Barbier, G., Liu, H.: Promoting coordination for disaster relief – from crowdsourcing to coordination. In: Salerno, J., Yang, S.J., Nau, D., Chai, S.-K. (eds.) SBP 2011. LNCS, vol. 6589, pp. 197–204. Springer, Heidelberg (2011)
8. Lughofer, E.: Hybrid Active Learning for Reducing the Annotation Effort of Operators in Classification Systems. Pattern Recognition **45**, 884–896 (2012)
9. Munro, R., Gunasekara, L., Nevins, S., Polepeddi, L., Rosen, E.: Tracking epidemics with natural language processing and crowdsourcing. In: 2012 American Association for Artificial Intelligence Spring Symposium, Toronto, Ontario, Canada (2012)
10. Palmer, G., D'Orazio, V., Kenwick, M., Lane, M.: The MID4 Data Set, 2002–2010: Procedures, Coding rules, and Description. Conflict Management and Peace Science (Forthcoming, 2015)
11. Ramirez-Loaiza, M.E., Culotta, A., Bilgic, M.: Anytime active learning. In: 28th American Association for Artificial Intelligence (2014)
12. Salek, M., Bachrach, Y., Key, P.: Hotspotting - a probabilistic graphical model for image object localization through crowdsourcing. In: DesJardins, M., Littman, M.L. (eds.) Proc. 27th American Association for Artificial Intelligence, July 14-18, Bellevue, Washington, USA. AAAI Press (2013)
13. Sheshadri, A., Lease, M.: Square: A benchmark for research on computing crowd consensus. In: First AAAI Conference on Human Computation and Crowdsourcing (2013)
14. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast-but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics (2008)
15. Von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: Using hard AI problems for security. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 294–311. Springer, Heidelberg (2003)