Modeling Complex Annotations

Alexander Braylan
Department of Computer Science
University of Texas at Austin
braylan@cs.utexas.edu

Matthew Lease
School of Information
University of Texas at Austin
ml@utexas.edu

Abstract

Modeling annotators and their labels is useful for ensuring data quality. However, while many models have been proposed to handle binary or categorical labels, prior methods do not generalize to *complex* annotation tasks (e.g., open-ended text, multivariate, structured responses) without devising new models for each specific task. To obviate the need for task-specific modeling, we propose to model distances between labels, rather than the labels themselves. Our methods are agnostic as to the distance function; we leave it to the annotation task *requester* to specify an appropriate distance function for their task. We propose three methods, including a Bayesian hierarchical extension of *multidimensional scaling*.

1 Motivation

Annotations provide the basis for supervised learning and evaluation. Given the importance of annotation, much work has considered models and measures of annotator behavior and labels (Dawid and Skene 1979; Smyth et al. 1995; Artstein and Poesio 2008; Passonneau and Carpenter 2014). The advent of inexpert crowd annotation (Snow et al. 2008) has stimulated a surge of further modeling work motivated by quality assurance with inexpert annotators. However, nearly all existing annotation models assume relatively simple labeling tasks, such as classification or rating.

Not all annotation tasks are so simple. Some tasks involve open-ended answer spaces (e.g., translation, transcription, extraction, generation) (Bernstein et al. 2010; Zaidan and Callison-Burch 2011; Li et al. 2016) or structured responses (e.g., annotating linguistic syntax or co-reference) (Paun et al. 2018). As methods for effective crowdsourcing continue to advance, we are seeing increasingly involved tasks, such as annotating lists or sequences (Nguyen et al. 2017), open-ended answers to math problems (Lin, Mausam, and Weld 2012), or even drawings (Ha and Eck 2017). Lacking task-independent, general-purpose models supporting aggregation for such tasks, aggregation is usually performed by task-specific models or by relying on additional human computation. Our research goal is to provide a general aggregation model supporting diverse complex annotation tasks.

We define *complex annotations* as any kind of annotation that could not be easily represented as a categorical variable or single-dimensional ordinal variable. Such tasks often involve a very large or infinite answer space, such that annotators are far less likely produce identical labels for the same item. For example, there can be multiple acceptable ways to translate a sentence (and even more incorrect ways). It thus makes less sense to assume a hard 0/1 loss in assessing annotator labels, but rather varying similarity to each other over the space of their possible values. Any method for aggregating complex annotations should be able to handle large and small dissimilarity as well as exact equality between them.

2 Background

When annotations are simple categorical variables, there is a rich literature of general-purpose and task-independent methods for aggregation. The most well-known model is from Dawid and Skene (1979), who provide an unsupervised or potentially semi-supervised method for inferring truth from user and item identifiers and labels. This probabilistic model can be trained via expectation maximization or Bayesian methods (Carpenter 2011). The Dawid-Skene model learns confusion matrices for each user representing that user's probability of giving the observed categorical label given the unknown true value. An alternative is to learn each user's probability of providing a correct answer (Demartini, Difallah, and Cudré-Mauroux 2012). Gold labels are not required, as parameters are learned through consensus between users. Dawid-Skene can be thought of as weighted voting, and it tends to outperform simple majority voting (Sheshadri and Lease 2013).

A common characteristic of aggregation methods for simple annotations is that they make use of statistical models to explain the collected data. Statistical models for crowd annotations provide a framework for several useful tools, including parameter inference, semi-supervised learning, and probabilistic task management. For tasks that collect complex annotations from the crowd, formulating statistical models can be very difficult because of the non-categorical likelihood functions. Designing such models requires both familiarity with the task domain and skill with mathematics and statistics. Some examples are a model based on Hid-

den Markov Models (HMM) that was developed to aggregate crowd-annotated sequences of text within documents (Nguyen et al. 2017) and a Chinese Restaurant Process (CRP) model for short free-response answers (Lin, Mausam, and Weld 2012). HMMs can only be used for data with time-dependence, and the CRP approach works when there are single discrete correct answers but not when there are continuous spaces of similarly correct ones. So far, no model has been proven effective for diverse complex annotation tasks.

3 Proposed Methodology

The key research question is how to provide a general framework for modeling complex crowd annotations without the need for task-specific statistical models. Our proposed idea for circumventing the need for task-specific models is to instead depend on task-specific *distance functions*, which are easier to reason about and exist in abundance for most of the tasks we might consider. As long as there exists an evaluation metric for comparing predictions to gold, that same metric could be used as a distance function.

Once a distance function is selected, the next step is to produce a distance dataset from the original dataset containing distances D_{iuv} for users $u \in U$ and $v \in U$ and items $i \in I$. This step can be done for each item in parallel by using the distance function to produce a matrix of annotation distances between all users that have annotated that item.

The distance dataset can be used to train a crowd annotation distance model. This model should infer true values for each item and might also infer helpful parameters describing user error and item difficulty. By modeling distances, we can now define the likelihood for a continuous variable (distances) rather than for complex objects (annotations). With both observed and inferred variables now entirely in continuous space, we avoid the main difficulty in designing statistical models for complex annotations.

Smallest Average Distance (SAD) Our first and simplest method operates local to each item, akin to majority voting. It selects the annotator label $\hat{L}_i$ for each item i with least average distance to all other labels for it.

Best Available User (BAU) Our second method selects the best annotation $\hat{L}_i$ for each item i by choosing the annotation from the most trusted annotator u'_i. Trust of u'_i is estimated as their average distance over the full dataset.

Multidimensional Annotation Scaling (MAS) Our final proposed method for modeling crowd annotation distances is inspired by Dawid-Skene and intended as a generalization of BAU and SAD that balances the contributions of each. The idea is to model a K-dimensional representation space in which the central point is taken as the estimated true item value, and annotation embeddings are estimated around that central point at norms regularized by expected user error.

In order to compute such annotation embeddings, we devise a statistical model based on *multidimensional scaling*. Multidimensional scaling is a method for estimating coordinates x of points given only a matrix of distances between those points by minimizing an objective function, generally $\sum(\|x_i - x_j\| - D_{ij})^2$. MAS utilizes the multidimensional scaling objective function in which the estimated coordinates serve as annotation embeddings. Instead of the data populating a single distance matrix, separate distance matrices correspond to each item. Because each user may annotate several items, the full dataset can be leveraged to compute *global* parameters representing user ability and serving as priors for the *local* parameters of each item's multidimensional scaling model. The resulting model is most easily expressed a hierarchical Bayesian model with a multidimensional scaling likelihood function. We specify this model in the Stan probabilistic programming language, which is equipped with algorithms for maximum a posteriori (MAP) estimation, variational inference (VI), and Markov chain Monte Carlo (MCMC).

4 Experiments

Our methods seek to support modeling and aggregation for datasets satisfying three conditions: complex labels, workers associated with identifiers, and gold labels to evaluate inferences. Several public datasets exist meeting two of those conditions, but meeting all three is rare. So far, we have conducted preliminary experiments on two real datasets meeting these conditions as well as two synthetic datasets. The largest real dataset we explore is a collection of 5,000 medical paper abstracts annotated by Amazon Mechanical Turk workers in (Nguyen et al. 2017). In this *Biomedical Information Extraction (IE)* task, workers annotate text spans describing populations enrolled in clinical trials. The other real dataset available to us is a collection of Urdu-to-English translations made by non-professional translators (Zaidan and Callison-Burch 2011). Of this set, we can only use the 300 items that have more than one annotation. Additionally, we develop a simulator for two possible kinds of complex annotations: syntactic parse trees and ranked lists. Syntactic parsing is a particularly interesting example for which corpora are challenging to produce even with trained linguists.

Our preliminary experiments compared our proposed methods against several baselines including selection of a **random user**'s annotation (RU) and the use of an **oracle** (OR) that selects the performance-maximizing annotations. For the sequences dataset, we additionally compare against the proposed method (HMM) and simplest baseline (MV) from (Nguyen et al. 2017). We do not compare to traditional aggregation models that require categorical annotations because the lack of exact matches between complex annotations suggests they would perform comparably to RU.

Table 1 displays the results of the experiments. The evalu-

	RU	BAU	SAD	MAS	Oracle
Parse Trees	0.879	<u>0.908</u>	0.905	**0.929**	0.965
Rankings	0.647	0.660	**0.679**	<u>0.673</u>	0.704
Sequences	0.549	<u>0.660</u>	0.658	**0.691**	0.824
Translations	0.254	0.259	<u>0.260</u>	**0.275**	0.302

Table 1: Preliminary experimental results. For sequences, MV achieves 0.647 and HMM achieves 0.697.

ation measures for the parse trees, rankings, sequences, and translations experiments are EVALB, Kendall's tau, F1, and BLEU, respectively. These also correspond to the distance functions used for each dataset.

These experiments provide some evidence that multidimensional annotation scaling (MAS) is a powerful general technique for aggregating complex labels. In the real datasets, it outperforms the other general aggregation methods by a wide margin. One interesting result is that the weaker methods are barely useful for the translation task, while MAS performs halfway between random and oracle. Having a powerful method for aggregating translation data might add substantial value to the collection of multiple translations. As for the simulated datasets, MAS outperforms the other methods significantly for the parsing task and very slightly underperforms SAD for the rankings task. Its underperformance in the rankings task might be comparable to how majority vote occasionally outperforms Dawid-Skene in simple tasks. Overall, MAS appears to succeed at its goal of combining the benefits of BAU and SAD in an approach that is more dependable and interpretable.

5 Challenges

The greatest challenge we face is finding appropriate data for our experiments. While there are plenty of datasets with complex labels, it is very rare to find such a dataset where multiple annotations were collected for each item and the annotator ID is preserved in the data. Without redundancy in annotations there is no aggregation to perform, and without annotator ID there is no way to infer annotator ability. In addition to these requirements, the dataset would need enough gold annotations to yield reliable testing results. In order to demonstrate that our methods generalize well to multiple unrelated tasks, we need several diverse complex annotation datasets that meet the above requirements.

Another challenge is to provide a more complete theoretical framing to our methods. We believe that a model of annotation distances such as MAS should, under varying configurations, generalize a wide range of crowd annotation models, including many of the commonly used models for simple label aggregation. A related goal is to provide a methodology for converting arbitrary models for simple annotations into distance-based models that behave identically on the simple task but can also be used for complex annotations.

Ultimately, we want to address the larger question of how to facilitate complex annotation collection. Aggregation modeling is one of the many tools for supporting this objective, and we hope to investigate how to adapt other methods from the crowdsourcing literature as well as novel approaches to challenges specific to complex tasks.

Acknowledgments

We thank the reviewers for their feedback on our submission and the crowd workers for the data they contributed for this research study. This was was supported in part by National Science Foundation grant No. 1253413. Any opinions, findings, and conclusions or recommendations expressed by the authors are entirely their own and do not represent those of the sponsoring agencies.

References

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4):555–596.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

Carpenter, B. 2011. A hierarchical bayesian model of crowdsourced relevance coding. In *TREC*.

Dawid, A. P., and Skene, A. M. 1979. 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.

Demartini, G.; Difallah, D. E.; and Cudré-Mauroux, P. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, 469–478. ACM.

Ha, D., and Eck, D. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Li, Y.; Song, Y.; Cao, L.; Tetreault, J.; Goldberg, L.; Jaimes, A.; and Luo, J. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4641–4650.

Lin, C. H.; Mausam, M.; and Weld, D. S. 2012. Crowdsourcing control: Moving beyond multiple choice. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Nguyen, A. T.; Wallace, B. C.; Li, J. J.; Nenkova, A.; and Lease, M. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, 299. NIH Public Access.

Passonneau, R. J., and Carpenter, B. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics* 2:311–326.

Paun, S.; Chamberlain, J.; Kruschwitz, U.; Yu, J.; and Poesio, M. 2018. A probabilistic annotation model for crowdsourcing coreference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1926–1937.

Sheshadri, A., and Lease, M. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, 156–164.

Smyth, P.; Fayyad, U. M.; Burl, M. C.; Perona, P.; and Baldi, P. 1995. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, 1085–1092.

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 *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. Association for Computational Linguistics.

Zaidan, O. F., and Callison-Burch, C. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1220–1229. Association for Computational Linguistics.