Distance-based Consensus Modeling for Complex Annotations

Anonymous EMNLP-IJCNLP submission

Abstract

Modeling annotators and their labels is useful for ensuring data quality. Though many models exist for 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 method, a Bayesian hierarchical extension of *multidimensional scaling*, is agnostic as to the distance function; we leave it to the annotation task *requester* to specify an appropriate distance function for their task. Evaluation shows the generality and effectiveness of the model across two complex annotation tasks: multiple sequence labeling and syntactic parsing.

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; 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 et al., 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 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). Methods for assessing and aggregating complex annotations ought to be flexible enough to model relative label similarity between labels, beyond simple exact match.

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 model, ZenCrowd, learns each user's prob-

ability of providing a correct answer (Demartini et al., 2012). These methods can be thought of as weighted voting, and they tend to outperform simple majority voting (Zheng et al., 2017) Gold labels are not required, as parameters are learned through consensus between users.

A common characteristic of aggregation methods for simple annotations is that they make use of probabilistic models to explain the collected data. Probabilistic 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 probabilistic models can be very difficult because the likelihood functions are not simple probability distributions. Designing such models requires both familiarity with the task domain and skill with mathematics and statistics. Some examples are a model with a Hidden Markov Model (HMM) likelihood function 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 et al., 2012). Both of these examples are limited in their applicability to the type of data they model (time-dependent and discrete variables, respectively). So far, no model has been proven effective for diverse complex annotation tasks. Our goal is to provide a more flexible option for complex tasks: a general-purpose and task-independent probabilistic model for aggregating complex annotations.

3 Multidimensional Annotation Scaling

The key idea to obviate the need for task-specific models is to instead rely on task-specific *distance functions*, which are easier to obtain and already exist for most annotation tasks of interest. 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 annotation dataset, containing distances D_{iuv} for users $u, v \in U$ and items $i \in I$. This distance dataset can be used to train a crowd annotation distance model. This model should grade the quality of annotations 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 probabilistic models for complex annotations.

Our proposed method for modeling crowd annotation distances is inspired by Dawid-Skene and intended to generalize a wide variety of aggregation models. 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 annotation embeddings, we devise a probabilistic model based on *multidimensional scaling* (Mead, 1992). Multidimensional scaling is a method for estimating coordinates ϵ of points given only a matrix of distances between those points by minimizing an objective function, generally $\sum(\|\epsilon_i - \epsilon_j\| - D_{ij})^2$. The estimated coordinate vectors carry meaning not in their absolute direction or magnitude, but rather in their position relative to each other.

Our model, *multidimensional annotation scaling* (MAS), is a hierarchical Bayesian probabilistic model with a multidimensional scaling likelihood function, in which the estimated coordinates serve as annotation embeddings. Instead of the data populating a single distance matrix, each item has a separate annotation distance matrix. Also, because each user may annotate several items, we leverage the full dataset to compute *global* parameters representing annotator reliability, which serve as priors for the *local* parameters of each item's multidimensional scaling likelihood.

We define the MAS model in Equations (1)-(4) and illustrate the basic premise in Figure 1.

$$\hat{L}_i = L_{iu'_i}, \quad u'_i = \text{argmin}_{u \in U(i)} \|\epsilon_{iu}\| \qquad (1)$$

For each item, the model selects an annotation $\hat{L}_i$ as its true value estimator, whose embedding ϵ_{iu} has the smallest norm out of all the annotations made by the annotators $U(i)$ of that item. In the MAS model, the origin in embedding space is taken to represent the true value for an item, so the norm of ϵ is understood as its distance from the truth. Unlike standard multidimensional scaling where the magnitude of the coordinates need not carry meaning, in our MAS model the magni-

tude of the annotation embeddings represents their *quality*. This measure also naturally lends itself to assigning partial credit to annotations. MAS assumes the annotation embedding space is *isotropic* because it does not depend on direction and *unimodal* because there is a single optimal point.

$$D_{iuv} \sim \mathcal{N}(\|\epsilon_{iu} - \epsilon_{iv}\|, \sigma) \quad (2)$$

Equation 2 is the generalized multidimensional scaling objective function expressed as a probabilistic likelihood. Maximizing the normal likelihood with free scale parameter σ minimizes the square error between observed distances in the data and learned distances in the embedding space.

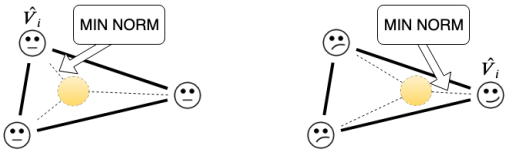
$$\epsilon_{iu} = \gamma_u \delta_i \frac{\tilde{\epsilon}_{iu}}{\|\tilde{\epsilon}_{iu}\|}, \gamma_u, \delta_i \in \mathbb{R}_+, \epsilon_{iu}, \tilde{\epsilon}_{iu} \in \mathbb{R}^K \quad (3)$$

The annotation embeddings ϵ comprise normalized raw coordinates $\tilde{\epsilon}$ as well as scale parameters γ representing user error and δ representing item difficulty. Normalizing the raw coordinates forces the scale parameters to entirely determine the embeddings' magnitudes. The model prefers to fit larger values of the scale parameters when those users and items are associated with larger distances in the data. When many annotations have small distances between each other, the model favors placing them closer to the origin compared to isolated annotations with higher distances from the others, thereby rewarding consensus. The model also favors placing annotations made by smaller-γ users closer to the center, thereby rewarding annotator reliability.

$$\log \gamma_u \sim \mathcal{N}(\log \bar{\gamma}, \Phi), \log \delta_i \sim \mathcal{N}(\log \bar{\delta}, \Psi) \quad (4)$$

The parameters γ and δ are given hierarchical Bayesian priors with global location parameters $\bar{\gamma}$ and $\bar{\delta}$ and with configurable scales Φ and Ψ, respectively, which are set to 1 by default. The use of hierarchical Bayesian modeling reduces arbitrary choices of hyperparameters by allowing global parameters to be learned empirically, and it has been adopted in much of the recent work in label aggregation (Carpenter, 2008; Raykar et al., 2010; Liu and Wang, 2012). Finally, we arbitrarily set the last hyperparameter $K = 8$ (untuned), slightly more than a typical five annotations per item.

For estimating the parameters of MAS, we specify the model in the Stan probabilistic programming language (Carpenter et al., 2017). Stan is equipped with algorithms for computing maximum a posteriori (MAP) estimates, variational inference, and Markov chain Monte Carlo. Our experiments utilize the fastest method, MAP.



(a) Equal-γ annotators (b) Different-γ annotators

Figure 1: Illustration of an item modeled by *multidimensional annotation scaling* (MAS). The emoji faces represent annotator labels, bold lines are observed distances between annotations, the golden circles are inferred true values, and dotted lines show the inferred magnitude of error for each annotation. When equal γ are learned for all annotators, the inferred true value is the geometric center. When different γ values are learned, the inferred true value is pulled closer to the more trusted annotators' labels. γ are learned from the observed distances over other items not displayed.

4 Experiments

Our methods support modeling and aggregation for datasets having 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 experiments on a real dataset meeting these conditions and a synthetic dataset whose configurations we vary over several experiments.

Real Annotations: Sequences. 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.

Synthetic Annotations: Parse Trees. Our parse tree simulator exploits a set of automatic parsers such as BLLIP (McClosky et al., 2006), MaltParser (Nivre et al., 2007), and the Stanford Parser, (Manning et al., 2014) implemented in NLTK (Loper and Bird, 2002) to produce multiple candidate parses of varying quality for a set of sentences drawn randomly from the Brown corpus. For each item, these parses are ordered by decreasing quality as measured by their *EVALB* score (Sekine and Collins, 1997). Simulated workers are randomly assigned skill parameters $\in [0, 1]$. They

Experiment					Baselines				Distance	
Dataset	Distance	$\|\|U\|\|$	$\|\|I\|\|$	$\|\|L\|\|$	RU	ZC	TMV	CHMM	MAS	Oracle
Parse Trees	EVALB	30	50	300	0.879	0.877	-	-	**0.929**	0.965
		10	50	300	0.873	0.876	-	-	**0.913**	0.975
		30	25	150	0.867	0.872	-	-	**0.889**	0.959
		10	25	150	0.866	0.865	-	-	**0.896**	0.976
Sequences	F1	91	191	1165	0.563	0.558	0.647	**0.697**	0.691	0.824

Table 1: Results. Metrics vary by task; larger is always better. The best result in each row is **bolded**. Lesser results whose difference is not statistically significant at the 0.05 level are indicated by underlining. The number of users $||U||$, number of workers $||I||$, and number of annotations $||L||$ vary.

are each given a random set of sentences, and their annotations are stochastically selected from the ordered list of candidate parses according to a geometric probability distribution, so that higher-skill workers tend to produce higher-quality parses.

Baselines. Our experiments compare our proposed method 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. We also compare against **ZenCrowd**, which models each worker's probability of providing correct labels and effectively performs weighted voting. For the sequences dataset, we additionally compare against the proposed Crowd-HMM method (CHMM) and simplest baseline token-wise majority vote (TMV) from Nguyen et al. (2017).

Results and Discussion. Table 1 displays results. In all experiments, MAS outperforms the other general aggregation methods by a wide margin. For the sequence annotation task, MAS outperforms TMV, which itself is still much better than RU. The CHMM probabilistic model specific to sequence annotation achieved the highest score, but there was no statistically significant difference vs. our task-agnostic MAS model. In principle, customized models for specific tasks should perform better than general-purpose alternatives, but at the cost of greater complexity and additional time and expertise to design. Benchmarking studies on aggregation for simpler annotation tasks (Zheng et al., 2017) have also shown that such off-the-shelf solutions are often remarkably competitive in practice vs. more customized models.

The question for requesters, then, is how much added benefit a custom model may deliver as return-on-investment vs. using an off-the-shelf, task-agnostic model such as MAS? It is also worth framing MAS in the context of two extremes: the cheapest option – only collecting one annotation per item (i.e., RU) – and the most expensive option – designing a custom probabilistic annotation model (e.g., CHMM) or custom human computation workflow for each new annotation task.

Across evaluation tasks, the learned ZC model performs nearly identically to RU. The likely culprit for ZC's lackluster performance is the large label space of complex annotation (Section 1), leading to poor annotator accuracy estimates for weighted voting. While we evaluate ZC specifically, its results are likely indicative of a larger family of existing, similar annotation models which estimate annotator reliability based on exact match between labels (Zheng et al., 2017).

5 Conclusion

Our MAS method bypasses the challenge of having to define task-specific probabilistic models for each new type of complex annotation by instead modeling the distances between annotations. Results on two complex tasks – sequence labeling and syntactic parsing – show improvement over general baselines and comparable performance to a task-specific probabilistic model for the sequence task. MAS thus appears to be useful, both for practical adoption and as a baseline against which new, bespoke annotation models for complex annotation tasks can be benchmarked.

One idea for future work is to extend our model to support complex tasks without assuming the annotation space is isotropic and unimodal (Section 3). This could extend MAS beyond *objective* tasks to also support *subjective* tasks (Tian and Zhu, 2012) having a space of valid responses which is wider and more uneven. Embedding scores produced by MAS could be used to identify a set of valid labels rather than a single best annotation.

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

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

Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322. ACM.

Bob Carpenter. 2008. Multilevel bayesian models of categorical data annotation. *Unpublished manuscript*, 17(122):45–50.

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

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478. ACM.

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

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650.

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

Chao Liu and Yi-Min Wang. 2012. Truelabel+ confusions: a spectrum of probabilistic models in analyzing multiple ratings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 17–24. Omnipress.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

Al Mead. 1992. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1):27–39.

An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, page 299. NIH Public Access.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

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

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio. 2018. A probabilistic annotation model for crowdsourcing coreference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1937.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

Satoshi Sekine and Michael Collins. 1997. Evalb: a bracket scoring program.

Padhraic Smyth, Usama M Fayyad, Michael C Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, pages 1085–1092.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. 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*, pages 254–263. Association for Computational Linguistics.

Yuandong Tian and Jun Zhu. 2012. Learning from crowds in the presence of schools of thought. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–234. ACM.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552.