Aggregating and Learning from Multiple Annotators

https://sites.google.com/view/alma-tutorial

Silviu Paun
Queen Mary University of London
s.paun@qmul.ac.uk

Edwin Simpson
University of Bristol
edwin.simpson@bristol.ac.uk

Abstract

The success of NLP research is founded on high-quality annotated datasets, which are usually obtained from multiple expert annotators or crowd workers. The standard practice to training machine learning models is to first adjudicate the disagreements and then perform the training. To this end, there has been a lot of work on aggregating annotations, particularly for classification tasks. However, many other tasks, particularly in NLP, have unique characteristics not considered by standard models of annotation, e.g., label interdependencies in sequence labelling tasks, unrestricted labels for anaphoric annotation, or preference labels for ranking texts. In recent years, researchers have picked up on this and are covering the gap. A first objective of this tutorial is to connect NLP researchers with state-of-the-art aggregation models for a diverse set of canonical language annotation tasks. There is also a growing body of recent work arguing that following the convention and training with adjudicated labels ignores any uncertainty the labellers had in their classifications, which results in models with poorer generalisation capabilities. Therefore, a second objective of this tutorial is to teach NLP workers how they can augment their (deep) neural models to learn from data with multiple interpretations.

1 Description

The disagreement between annotators stems from ambiguous or subjective annotation tasks as well as annotator errors. Crowdsourcing with non-expert annotators is especially prone to annotation errors, sometimes caused by workers who do not attempt to provide correct annotations (spammers). The traditional resolution to this problem is redundant labeling: collect multiple interpretations from distinct coders, allowing the resource creators to later aggregate these labels. To this end, probabilistic models of annotation have been successfully used to learn the coders' behavior and distill the labels from noise.

The research on models of annotation contains a large body of work spanning multiple decades (from the work on latent structure analysis back in the early 70s), and has been substantially debated over the years at dedicated conferences such as HCOMP and workshops, e.g., from The People's Web Meets NLP (Gurevych and Zesch, 2009), to CrowdML (http://crowdml.cc/), and more recently AnnoNLP (Paun and Hovy, 2019). The plethora of models that had been published even prompted some researchers to ask, challengingly, whether the problem of aggregating crowd labels had been solved (Zheng et al., 2017). As anticipated, there are still unaddressed issues – in particular, the bulk of work has focused on classification tasks, leaving room for innovation in other areas. The NLP field specifically contains a number of tasks with unique characteristics not considered by standard models of annotation. For example, in sequence labeling tasks such as part of speech tagging or named entity recognition, nearby labels have known inter dependencies. In other tasks such as anaphoric annotation for coreference resolution, the coders are asked to provide labels that are not from a fixed set of categories but consist of textual mentions. Another example is pairwise preference labelling, where coders are asked to choose the instance from a pair that most strongly reflects a quality of interest, such as relevance to a topic or convincingness of an argument, with the goal of inferring an overall ranking of text instances. Researchers have observed these gaps in the literature and are addressing them. A key objective of this tutorial is to connect NLP researchers with state-of-the-art aggregation methods suitable for canonical NLP tasks, covering classifications (Yan et al., 2014), sequence labels (Nguyen et al., 2017;

Proceedings of EACL: Tutorials, pages 6–9
April 19 - 20, 2021. ©2020 Association for Computational Linguistics

Simpson and Gurevych, 2019), anaphoric interpretations (Paun et al., 2018b) and pairwise preference labels (Simpson and Gurevych, 2020).

Resource creators can use aggregation methods to adjudicate the disagreements inherent in annotated data, but at times, when the resource is to serve as training data to a machine learning model, the noise distillation procedure does not have to be separated and can be integrated into the learning process. In fact, by following the convention and training with adjudicated labels we ignore any of the uncertainty the labellers had in their classifications. Including the coders' disagreements in the learning signal offers the models a richer source of information compared to adjudicated labels: they include not only the consensus, but may also indicate ambiguity, and how the humans make mistakes. This improves the generalisation capability of the models and offers them a more graceful degradation with less ridiculous mistakes (Peterson et al., 2019; Guan et al., 2018). Some of these approaches can also be used for their noise distillation capabilities, as their learning processes also produce aggregated labels that leverage not only coder annotation patterns but also the knowledge of the task accumulated by the model (Cao et al., 2018; Rodrigues and Pereira, 2018; Albarqouni et al., 2016; Chu et al., 2020). Often, this means that fewer redundant labels are required to attain the desired level of accuracy for the aggregated labels. Thus, a second objective of the tutorial is to teach NLP researchers how they can augment their existing (deep) neural architectures to learn from data with disagreements.

1.1 Learning outcomes

We aim to guide NLP practitioners through the emerging body of literature and train them to:

1. Apply aggregation methods and interpret their output predictions;

2. Identify state-of-the-art aggregation methods for canonical NLP tasks: classification, sequence labelling, anaphoric interpretations, and pairwise preferences;

3. Augment a (deep) neural network architecture to learn from data with multiple interpretations.

1.2 Type of tutorial

Introductory. The content will reference and explain well-established work but the focus is on novel, state-of-the-art methods.

2 Outline of Tutorial

Part 1. Motivation and Early Approaches to Annotation Analysis

1. Introduction to the field. Shortcomings of early practices.

2. Modeling the annotation process with a probabilistic model. How to encode our assumptions about the coders, the difficulty of the items, and their interactions. Using hierarchical models to alleviate sparsity.

Part 2. Advanced Models of Annotation

3. Aggregating sequence labels. In such tasks the labels of nearby items have known interdependencies. We discuss probabilistic approaches that model these sequential dependencies both between the ground truth labels and the annotations. We exemplify the utility of the methods on a NER task.

4. Aggregating anaphoric judgements for coreference resolution. For this task the annotation scheme does not use a fixed class space. The judgements here consist of labels assigned to textual mentions that mark when new entities are introduced into the discourse, non refering expressions such as expletives or predicative NPs, and recent antecedents of previously discussed entities. We explain how to apply a probabilistic mention-pair model to aggregate the labels and build coreference chains.

5. Preference labels: why comparisons can be more reliable than ratings or classifications. We show how to reformulate NLP tasks with ambiguous categories or scores as preference learning, giving an example applications related to argument persuasiveness. We introduce probabilistic approaches for aggregating preference judgements to infer a gold standard ranking.

6. Aggregation with Variational Autoencoders. This framework allows us to use neural networks to capture complex non linear relationships between the annotations and the ground truth. By doing so, we avoid having to manually identify and specify these relationships as in standard probabilistic models.

Part 3. Learning with Multiple Annotators

7. Learning with human uncertainty. The standard for training classifiers is to learn from data where each example has a single label. In doing so however any uncertainty the labellers had in their classification is ignored. We discuss here a few approaches to learning from the label distributions produced by the coders, which can improve classifier performance.

8. Humans are noisy. The success of the approaches from the previous point relies on the quality of the target distributions, i.e., whether the collected annotations offer a good representation of the coders' dissent. That may not always be the case, e.g., when their number is too low to get a good proxy for the human uncertainty, or when noise intervenes and skews the distributions. For this purpose we discuss a few training approaches that also capture the accuracy and alleviate the bias of the coders, with an emphasis on neural methods.

Part 4. Practical Session

9. Introduce the audience to an implementation of a probabilistic (Dawid and Skene, 1979) and a neural (Rodrigues and Pereira, 2018) model of annotation. The instructors will provide an example dataset and implementations of the two models then run through a few short exercises that will help the audience to understand and apply the methods to a real NLP task. The exercises will include comparing majority voting with the model of Dawid and Skene (1979) and training a downstream model on adjudicated labels compared to training directly on crowdsourced labels with (Rodrigues and Pereira, 2018). The dataset and code will be provided freely on the tutorial website.

2.1 Audience prerequisites

The audience may benefit from basic knowledge of probability theory, and of neural networks, but all concepts will be introduced from scratch. For the exercises, basic programming skills of Python and familiarity with Keras (in Tensorflow) are useful. The NLP task examples do not require detailed knowledge of the tasks themselves and the course is designed to be accessible for researchers who are new to the field.

2.2 Recommended reading list

Recommendations for part 1:

1. Passonneau and Carpenter (2014)
2. Paun et al. (2018a)

Recommendations for part 2:

3. Simpson et al. (2019)
4. Yin et al. (2017)

Recommendations for part 3:

5. Peterson et al. (2019)
6. Rodrigues and Pereira (2018)

3 Presenters

Silviu Paun, Queen Mary University of London (s.paun@qmul.ac.uk).

Silviu (https://silviupaun.com/) is a postdoctoral researcher with expertise in label aggregation and training of models on data with disagreements. He is part of the DALI project (http://dali.eecs.qmul.ac.uk/), in charge of the analysis of the annotations collected using Phrase Detectives, a GWAP (game with a purpose) developed for gathering labels for coreference resolution, with over 5 million judgements collected. He is regularly involved in machine learning seminars, one of which he organises at Queen Marry University of London, delivering lectures on probabilistic models for NLP applications and parameter estimation techniques.

Edwin Simpson, University of Bristol (edwin.simpson@bristol.ac.uk).

Edwin is a lecturer (assistant professor) who is leading new courses on Dialogue and Narrative and Text Analytics at the University of Bristol. During his PhD, he researched probabilistic aggregation methods for crowdsourced data, working with Zooniverse (https://www.zooniverse.org/), the world's largest volunteer crowdsourcing effort, and has advised numerous partners in science and industry on crowdsourced data aggregation (e.g., https://alephinsights.com). Recently, he led a well-received seminar course and lectures on crowdsourcing and gave tutorials on Bayesian methods at Technische Universität Darmstadt. His research involves developing preference learning techniques for NLP tasks and learning from interactions with end users and crowds.

References

S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. 2016. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321.

Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. 2018. Max-mig: an information theoretic approach for joint learning from crowds. In *International Conference on Learning Representations*.

Zhendong Chu, Jing Ma, and Hongning Wang. 2020. Learning from crowds by modeling common confusions.

Alexander Philip Dawid and Allan M Skene. 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.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification.

Iryna Gurevych and Torsten Zesch, editors. 2009. *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*. Association for Computational Linguistics, Suntec, Singapore.

An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada. Association for Computational Linguistics.

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, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018a. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

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

Silviu Paun and Dirk Hovy, editors. 2019. *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*. Association for Computational Linguistics, Hong Kong.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Filipe Rodrigues and Francisco C Pereira. 2018. Deep learning from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

Edwin Simpson and Iryna Gurevych. 2020. Scalable Bayesian preference learning for crowds. *Machine Learning*, pages 1–30.

Edwin D. Simpson and Iryna Gurevych. 2019. A Bayesian approach for sequence tagging with crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1093–1104, Hong Kong, China. Association for Computational Linguistics.

Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. 2014. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327.

Li'ang Yin, Jianhua Han, Weinan Zhang, and Yong Yu. 2017. Aggregating crowd wisdoms with label-aware autoencoders. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1325–1331.

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