Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth

Mike Schaekermann
University of Waterloo
Waterloo, ON N2L 6E5
Canada
mschaeke@uwaterloo.ca

Alex C. Williams
University of Waterloo
Waterloo, ON N2L 6E5
Canada
Alex.Williams@uwaterloo.ca

Edith Law
University of Waterloo
Waterloo, ON N2L 6E5
Canada
Edith.Law@uwaterloo.ca

William Callaghan
University of Waterloo
Waterloo, ON N2L 6E5
Canada
William.Callaghan@uwaterloo.ca

CHI'16, May 7–May 12, 2016, San Jose, USA.

DOI string from ACM form confirmation

Abstract
In this position paper, we challenge the conventional assumption that, in supervised machine learning, ground truth data always needs to provide exactly one correct label per training example. Recognizing the fact that there are various circumstances under which domain experts may disagree in a classification task, we argue that expert disagreement does not necessarily always have to be noise, as traditional approaches hypothesize, but, instead, may as well be considered valuable signal. In particular, we emphasize that certain types of ambiguity in ground truth data may be inherently irresolvable, alleging examples from the fields of crowdsourcing in medicine and papyrology. As a solution, we propose a new hybrid framework for dealing with uncertainty in ground truth that fully acknowledges the notion of irresolvable ambiguity, and iteratively elicits feedback from crowdworkers to decide whether an instance of disagreement is resolvable or not in order to train the intended classifier accordingly.

Author Keywords
ground truth; ambiguity; expert disagreement; crowdsourcing; supervised learning.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

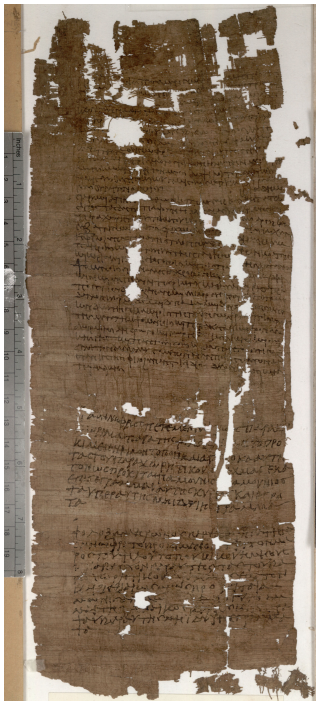

Figure 1: "Cession of Vacant Lot", P.Oxy 4586. This document is made-up of fragments from the Oxyrhynchus papyri collection, the largest known aggregation of deteriorated ancient Greek papyrus fragments in the world. We thank the Egyptian Exploration Society for providing access to this image.

Introduction

A common assumption in supervised machine learning is the idea that objects can be unambiguously classified into categories, and the quality of the ground truth data can be measured by the extent to which annotators agree with one another. In practice, many classification tasks are ambiguous, due to either missing information in the context (blurry image, broken fragments, etc) or the existence of multiple interpretation. To deal with ambiguity, most machine learning approaches advocate the *elimination* of ambiguity, by forcing the objects to belong to one of the categories through majority votes. Recent work [7, 2] argue that disagreement is not noise, but a meaningful source of information that can be used to filter out bad annotators or inherently ambiguous instances. In this work, we propose a machine learning framework that iteratively elicits feedback from crowdworkers to determine whether the ambiguity of a particular instance is resolvable or irresolvable, then train using these instances differently depending on the results of the crowd judgments.

Related Work

The phenomenon of disagreement between domain experts has been studied extensively from various perspectives across scientific disciplines [1, 5, 9, 10]. Although most work agree that heterogeneity among expert opinions may have different causes (like divergent understanding of the problem definition, unequal access to contextual information, etc.) [10], to this point, the majority of supervised machine learning approaches generalize that ambiguity is per se an indicator of noise in ground truth and should therefore be eliminated [11, 3]. More recent research from the field of crowdsourcing demonstrates that disagreement between expert annotators may instead be utilized to identify low-quality crowdworkers and "poor" training examples [7, 8, 2, 1].

Examples of Ambiguous Classification Tasks

Greek Letter Identification

Papyrology, the study of ancient literature and manuscripts written on papyrus, is one such field that inherently thrives on expert disagreement. The entirety of today's known papyrus collections, which date as far back as the 1st century, have been serendipitously discovered from archaeological excavations over the last 200 years. Despite being buried underground for hundreds of years, exposure to the harsh Egyptian climate has caused many of the documents to deteriorate and fragment, subsequently creating a plethora of small, partially-legible papyrus fragments. As such, one fundamental task of papyrology is transcribing a papyrus fragment.

Disagreeance in transcription among expert papyrologists is common, especially for cases where multiple fragments are being "joined" together and a gap of arbitrary length exists between the joined fragments (see Figure 1). By leveraging their mastery of the language and assessing the scribe's literacy, experts infer content for the gap and transcribe the surrounding text. The ambiguity of a gap's content often leads to many interpretations and conjectures of the same text, which in most cases can only be resolved after a new fragment or copy of a preexisting fragment has been found with a clear reading of the same text.

Sleep Stage Classification

Sleep staging is the classification of electroencephalographic (EEG) recordings into five stages. The analysis of sleep stages can be used to characterize a wide range of medical conditions, such as sleep disorders and epilepsy [4]. Figure 2 shows an example of an EEG signal which experts can classify sleep into Wake, N1, N2, N3, REM. Each sleep stage is characterized by the presence of certain EEG features; for example, sleep spindles (Figure 3)

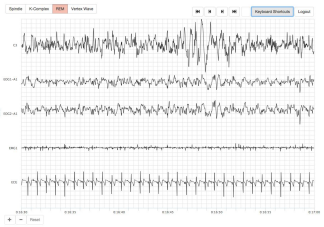

Figure 2: Sleep Stage Classification Task

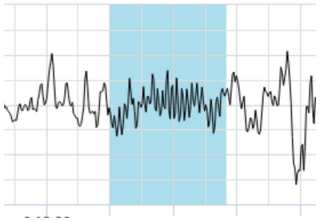

Figure 3: The sleep spindle is defined by the American Academy of Sleep Medicine as “a train of distinct waves with frequency 11-16Hz [...] with a duration of [greater than or equal to] 0.5 seconds, usually maximal in amplitude over the central region” [6].

are prevalent in N2 sleep. At the classification level, ambiguity can arise as one sleep stage transitions to another, where EEG features associated with both stages can be present.

At the same time, ambiguity can arise at the feature detection level (i.e., identification of sleep spindles) due to inexact morphological definitions (e.g., “a sleep spindle has a diamond or football shape” [11]). There are different scenarios of expert disagreement: (a) one expert can identify a region as a spindle where the other may not, (b) the annotation of one expert may overlap only partially with that of another expert. In the former case, the possibility that experts will reach unanimous consensus appears considerably more unlikely than in the latter case where experts agree on the presence of a sleep spindle within a certain temporal range. In both cases, however, machine learning algorithms would benefit from elicitation of feedback from expert annotators as to whether a unanimous agreement can or cannot be reached.

Proposed Human-Machine Framework for Learning From Ambiguous Ground Truth

We propose a new hybrid human-machine framework for learning from ambiguous ground truth, based on the key argument that disagreement among domain experts may either be (a) resolvable or (b) irresolvable (see Figure 4). Resolvable instances of ambiguity are defined as those cases for which the probability that experts will eventually reach unanimous consensus increases with increasing richness of contextual information. In contrast to that, for irresolvable cases of ambiguity, experts will not reach unanimity regardless of the amount of context, available for a given labeling task. In other words, irresolvable ambiguity arises in situations where human interpretation is subject not only to the availability of contextual information, but also to annotator-specific personal intrinsic features (e.g., experience, taste, mood) which might or might not change over time.

Harnessing this relationship, we propose an iterative learning framework in which cases of high disagreement between crowdworkers will be classified as either (a) resolvable or (b) irresolvable. To this end, more contextual information, related to the training example at hand, will be requested from and revealed to crowdworkers. Those training examples for which the level of disagreement decreases in response to the availability of new information will be classified as resolvable, whereas cases in which the ambiguity metrics are not responsive to the provision of new information will be considered irresolvable.

Based on this binary classification with respect to resolvability of ambiguous training examples, the machine learning algorithm will be trained separately. If new contextual information arises (e.g., new high-quality scans of papyri are uploaded after a recent excavation, or previously unknown medical conditions of a patient with sleep disorder are diagnosed), a new iteration cycle will be initiated to re-evaluate whether previously irresolvable cases of ambiguity can now reach a unanimous consensus among annotators. To account for such an update to the available gold standard data the classifier would be retrained accordingly.

Conclusion

In this paper, we highlighted the demand for a novel holistic approach towards supervised machine learning that fully embraces the notion of ambiguity in ground truth and distinguishes between resolvable and irresolvable disagreement between domain experts. To this end, we proposed an iterative feedback loop between the machine learning algorithm and a group of crowdworkers to investigate whether certain instances of high disagreement converge towards

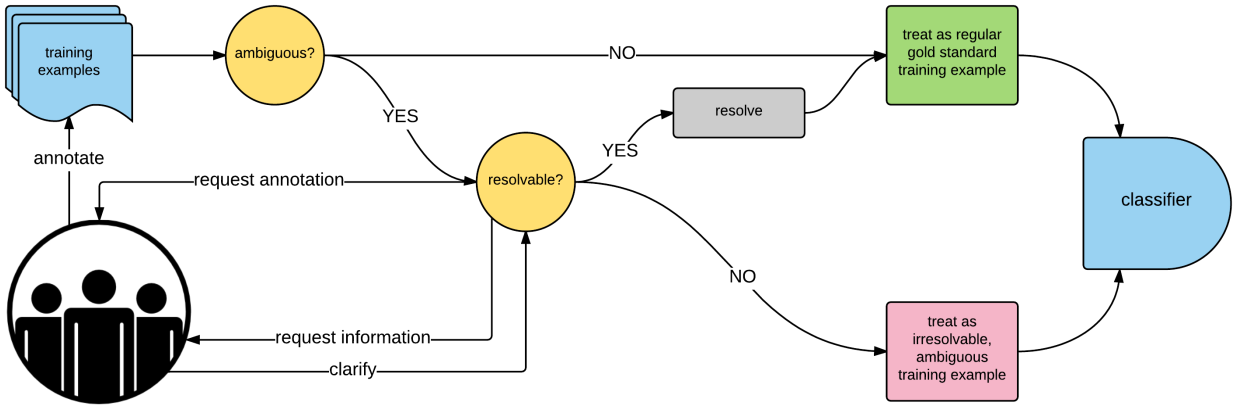

Figure 4: A top-level overview on the workflow for a hybrid human-machine framework for learning from ambiguous ground truth. Image of crowdworkers (bottem left): "group" by Justin Blake from the Noun Project.

unanimous consensus once more contextual information is provided to annotators. We argued that such a hybrid human-machine framework would cater more precisely to the requirements of various real-world domains, alleging examples from medicine and papyrology, while leaving concrete implementation details open for future research.

References

[1] Lora Aroyo and Chris Welty. 2013. *Measuring Crowd Truth for Medical Relation Extraction*. Technical Report. AAAI, Menlo Park, California. https://www.aaai.org/ocs/index.php/FSS/FSS13/paper/viewFile/7627/7543

[2] Lora Aroyo and Chris Welty. 2014. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (2014), 15–24.

[3] Arthur Carvalho and Kate Larson. 2013. A Consensual Linear Opinion Pool. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press, Beijing, China, 2518–2524. http://dl.acm.org/citation.cfm?id=2540128.2540491

[4] Elizabeth Ann Fehrmann. 2013. *Automated Sleep Classification Using the New Sleep Stage Standards*. Master Thesis. Rochester Institute of Technology. http://scholarworks.rit.edu/theses/3123/

[5] Luciana Garbayo. 2014. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making* 539 (2014), 35–45. http://link.springer.com/10.1007/978-3-319-04280-0_5

[6] Conrad Iber, Sonia Ancoli-Israel, Andrew L Cheeson Jr., and Stuart F Quan. 2007. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, Westchester, Illinois.

[7] Oana Inel, Lora Aroyo, Chris Welty, and Robert-Jan Sips. 2013. Domain-Independent Quality Measures for Crowd Truth Disagreement. In *The 12th International Semantic Web Conference (ISWC2013)*. http://data.semanticweb.org/workshop/derive/2013/proceedings/paper-01/html

[8] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *13th International Semantic Web Conference (ISCW2014)*. Springer Verlag, Cham, 486–504. DOI: http://dx.doi.org/10.1007/978-3-319-11915-1_31

[9] Elisabetta Lalumera. 2015. Overcoming Expert Disagreement In A Delphi Process. An Exercise In Reverse Epistemology. *HUMANA.MENTE Journal of Philosophical Studies* 28 (2015), 87–104. http://www.thehumanmind.eu/PDF/Issue28_Papers_Lalumera.pdf

[10] Jeryl L Mumpower and Thomas R Stewart. 1996. Expert Judgement and Expert Disagreement. *Thinking and Reasoning* 2, 23 (1996), 191–21.

[11] Simon C Warby, Sabrina L Wendt, Peter Welinder, Emil G S Munk, Oscar Carrillo, Helge B D Sorensen, Poul Jennum, Paul E Peppard, Pietro Perona, and Emmanuel Mignot. 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods* 11, 4 (feb 2014), 385–392. DOI: http://dx.doi.org/10.1038/nmeth.2855