Alessandro Checco
Gianluca Demartini
Ujwal Gadiraju
Cristina Sarasua
Lora Aroyo
Anca Dumitrache
Pavreen Paritosh
Alex Quinn
Chris Welty (Eds.)

Subjectivity, Ambiguity and Disagreement (SAD) in Crowdsourcing 2018

CrowdBias'18: Disentangling the Relation Between Crowdsourcing and Bias Management

Workshops co-located with HCOMP 2018

Zurich, Switzerland, July 5, 2018

Proceedings

Editors' addresses:
a.checco@sheffield.ac.uk, demartini@acm.org, gadiraju@l3s.de, sarasua@ifi.uzh.ch, lmaroyo@gmail.com, anca.dmtrch@gmail.com, pkp@google.com, aq@purdue.edu.

Preface

SAD 2018

For a summary of SAD workshop, we refer to http://www.sadworkshop.wordpress.com/.

CrowdBias'18

The CrowdBias'18 workshop (https://sites.google.com/view/crowdbias) was held on 5th July 2018 during the first day of the AAAI Conference on Human Computation at the University of Zurich, Switzerland. The goal of this workshop was to analyze both existing biases in crowdsourcing, and methods to manage bias via crowdsourcing. The workshop discussed different types of biases, measures and methods to track bias, as well as methodologies to prevent and mitigate bias.

Crowdsourcing has become a successful and widely used means to obtain human input on a large-scale. Human intelligence is needed to evaluate various systems, augment algorithms, perform high quality data management among a hoard of other applications. Humans though, have various cognitive biases that influence the way they interpret statements, make decisions and remember information. If we use crowdsourcing to generate ground truth, it is important to identify existing biases among crowdsourcing contributors and analyze the effects that their biases may produce and propagate. At the same time, having access to a potentially large number of people can give us the opportunity to manage the biases in existing data and systems.

Alessandro Checco, Gianluca Demartini, Ujwal Gadiraju and Cristina Sarasua served as co-chairs for this workshop. The workshop consisted of 2 keynote talks, 7 paper presentations and a moderated discussion. We provided a platform and framework for discussion among scholars, practitioners and other interested parties, including crowd workers, requesters and crowdsourcing platform managers. In the first keynote talk, Jahna Otterbacher (Open University Cyprus) discussed social biases in human-machine information systems and encouraged the audience to think of algorithmic transparency and accountability. In a second keynote, a long-time Turker, researcher and founder of Turker Nation, Kristy Milland talked about bias from the perspective of crowd workers. Kristy Milland stressed the fact that crowd workers' motivations may change in different contexts and over time, and emphasized the need to treat crowd workers cautiously and fairly.

Papers that were presented outlined the themes of aggregation bias, measuring how opinion bias influences crowdsourced labelling tasks, and pitting biases emanating from experts in comparison to the crowd. Other major themes included measuring bias in data or content using crowdsourcing, bias in task selection, sampling biases during recruitment, and biases induced due to the work environments. The moderated discussion resulted in identifying the need for a 'taxonomy of biases in crowdsourcing', possibly extending other classifications

of general biases on the Web, that can guide future work in understanding and managing bias in crowdsourcing. Investigating the various sources of bias, and developing methods to present bias related information to end users was identified as an important challenge in sociotechnical and crowdsourcing systems. The influence of task types in propagating and mitigating biases was discussed. Finally, the paid crowdsourcing paradigm was recognized as a special realm where additional factors such as worker motivation, self-selection, rewards, task design, etc. can influence task outcomes.

July 2018

Alessandro Checco, Gianluca Demartini,
Ujwal Gadiraju, Cristina Sarasua

Organizing Committee - SAD 2018

Lora Aroyo, Vrije Universiteit Amsterdam
Anca Dumitrache, Vrije Universiteit Amsterdam
Praveen Paritosh, Google
Alex Quinn, Purdue University
Chris Welty, Google

Organizing Committee - CrowdBias'18

Alessandro Checco, The University of Sheffield
Gianluca Demartini, University of Queensland
Ujwal Gadiraju, Leibniz Universität Hannover
Cristina Sarasua, University of Zurich

Program Committee - CrowdBias'18

Omar Alonso, Microsoft
Ricardo Baeza-Yates, NTENT
Jo Bates, The University of Sheffield
Michele Catasta, Stanford University
Alessandro Bozzon, Delft University of Technology
Paul Clough, University of Sheffield
Djellel Difallah, New York University
Carsten Eickhoff, Brown University
Leo Ferres, Universidad del Desarrollo
Rochelle Laplante, Professional Crowdworker
Kristy Milland, Turker Nation
Alexandra Olteanu, IBM
Bibek Paudel, University of Zurich
Mike Schaekermann, University of Waterloo

Contents

Crowdsourced Measure of News Articles Bias: Assessing Contributors' Reliability

Emmanuel Vincent[1], Maria Mestre[1]

[1] Factmata Ltd.,
114 Whitechapel High St,
London E1 7PT
{emmanuel.vincent, maria.mestre}@factmata.com

Abstract. We tackle the challenge of building a corpus of articles labelled for their political bias by relying on assessments provided by a crowd of contributors. The definition of 'bias' can be ambiguous to participants and both the targets of the ratings (articles) and the source of ratings (contributors) can be biased in some ways. In this paper, we explore techniques to mitigate this subjectivity and learn about the bias of both articles and contributors from the agreements and disagreements among their assessments. We report on the effectiveness of using a set of gold-standard articles to evaluate the reliability of contributors and discuss work in progress to evaluate the bias of contributors from their relative assessments of articles' bias.

1 Introduction

News providers are routinely accused of displaying political bias and this has become a pressing issue as the polarization is the population is increasing, notably in the US (Martin and Yurukoglu 2017). Social media platforms where users increasingly get their news have also been pointed as a source of increased polarization among the public and for favoring the rise of extremely biased information providers, whose inflammatory language is particularly prone to spreading on social media (Marwick and Lewis 2017).

Increased partisanship in news results in enhanced polarization in societies, which undermines democracy and is sometimes a factor in increasing ethnic violence (Minar and Naher 2018). For this reason, several governments have recently attempted to address this growing concern by developing legislation against "fake news". Advertisers are also increasingly interested in measures and detection of extreme bias in online content, as their brand values might be incompatible with funding hyper partisan or divisive content.

In this context, finding scalable ways to assess the bias of articles or information providers is a pressing challenge. This paper presents the first step of ongoing work in Factmata's effort to create a corpus of articles annotated for political bias relying on a crowdsourced approach and design of a system to identify the most reliable contributors.

2 Related Work

Several websites compile lists of news outlets characterized by their bias, one of the most prominent being Media Bias/Fact Check (MBFC)[1]. Like other similar initiatives, MBFC relies on the classification established by a few individuals and classifies news sources at the outlet level, based on analysis of a few articles published by the outlet. Approaches based on natural language processing (NLP) have been used to scale up bias detection, as Lazaridou and Krestel (2016), for example, who analyzed which politicians were being quoted by two major UK outlets, and showed this provided an indication of the outlets' political biases. Patankar and Bose (2016) have approached the challenge of determining bias at the individual news articles level using NPL tools that detect non-neutral sentence formulations based on Wikipedia non-NPOV corpus.

Further automation of bias detection based on Machine Learning approaches will need the creation of large datasets of labeled articles, and in this case crowdsourced solutions offer interesting scalability perspective. Budak, Goel and Rao (2016) performed a large-scale analysis of media bias in which contributors recruited on Mechanical Turk assessed the political bias of more than 10,000 articles from major media outlets covering US politics. However studies like this one did not investigate how to learn about the bias and reliability of contributors from their assessments. More insight can be learned in this respect from online rating systems, in relation to which research has been lead on trust and reputation to identify contributors' reliability and identify potentially biased or spam users (read Swamynathan, Almeroth and Zhao 2010 for an overview). The challenge is to develop a system that allows to learn both about the bias of the news articles that are being labeled and the bias and reliability of the contributors who provide the labels.

3 Data

3.1 Crowdsourcing bias assessments

We drew articles from a pilot study, representing a corpus of 1,000 articles on which ads had been displayed for the account of a customer; these thus form a sample of highly visited news articles from mainstream media as well as more partisan blog-like "news" sources. We used the Crowdflower platform[2] to present these articles to participants who were asked to read each article's webpage and answer the question: "Overall, how biased is this article?", providing one answer form the following five-point bias scale:

1. Unbiased
2. Fairly unbiased
3. Somewhat biased
4. Biased

[1] https://mediabiasfactcheck.com/
[2] https://crowdflower.com/

5. Extremely biased

To guide their assessments, we provided contributors with more details regarding how to classify articles in the form of a general definition of biased article as well as examples of articles with their expected classification (see Appendix 1 for details of the instructions). We chose a five-point scale to allow contributors to express their degree of certainty, leaving the central value on the scale (3) for when they are unsure about the article bias while the values 1 and 2 or 4 and 5 represent higher confidence that the article is respectively unbiased or biased to a more (1 and 5) or less (2 and 4) marked extent. Fifty participants contributed to the labeling and five to fifteen contributors assessed each article (see Appendix 2 for an example).

3.2 'Gold' dataset

To assess the reliability of contributors, we also asked two expert annotators (a journalist and a fact-checker) to estimate which bias ratings should be counted as acceptable for a quarter of all the articles in the dataset. For each article in this 'gold' dataset, the values provided by the two experts are merged. Two values are typically found to be acceptable for an article (most often 1 and 2, or 4 and 5), but sometimes three values are deemed acceptable and less often one value only: typically when both experts agree the article is either clearly extremely biased or not biased at all (e.g. because it covers a trivial and non-confrontational topic in the latter case). When experts disagree on the nature of the bias, providing a set of acceptable ratings as strictly greater than three for one and strictly lower than three for the other, the article is not considered in the gold dataset.

4 Analysis of results

4.1 Assessing contributors' reliability

As a first approach to guide us regarding the quality of data we collected, we performed a comparison of contributors' rating with the gold dataset ratings. Building on the "Beta reputation system" framework (Ismail and Josang 2002), we represent users' reliability in the form of a beta probability density function. The beta distribution $f(p|\alpha,\beta)$ can be expressed using the gamma function Γ as:

$$f(p|\alpha,\beta) = \Gamma(\alpha+\beta)/(\Gamma(\alpha).\Gamma(\beta)) \, . p^{\alpha}(1-p)^{\beta-1} \, . \quad (1)$$

where p is the probability a contributor will provide an acceptable rating, and α and β are the number of 'correct' (respectively 'incorrect') answers as compared to the gold. To account for the fact that not all incorrect answers are as far from the gold, we further weight the incorrect answers as follows: an incorrect answer is weighted by a factor of 1, 2, 5 or 10 respectively if its shortest distance to an acceptable answer is 1, 2, 3 or 4 respectively. So β is incremented by 10 (resp. 2) for a contributor providing a rating of 1 (resp. 4) while the gold is 5 (resp. 2) for example. We use the expectation

value of the beta distribution $R = \alpha/(\alpha + \beta)$ as a simple measure of the reliability of each contributor. See figure 1 for examples of reputation function obtained for (a) a user with few verified reviews, (b) a contributor of low reliability and (c) a user of high reliability.

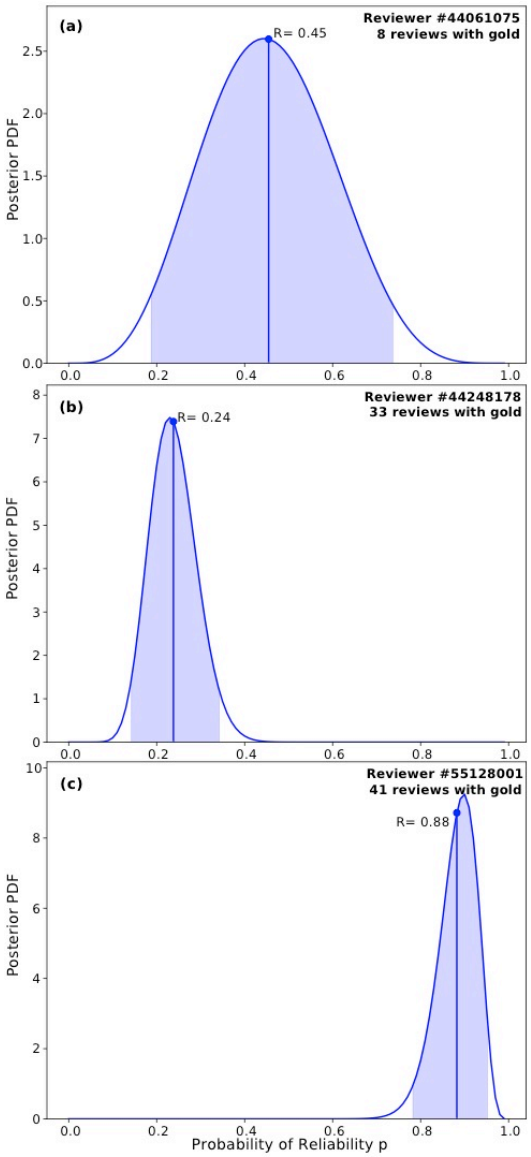


Fig. 1. Examples of reputation function obtained for (a) a user with few verified reviews for whom the uncertainty is still large, (b) a contributor of low reliability and (c) a user of high reliability. Shading shows the 95% probability interval.

Inter-rater reliability. We calculated Krippendorff's alpha to measure the inter-rater agreement (Krippendorff 2011). When we include every worker, we obtain a value for alpha of 0.078, which can be interpreted as a very low agreement. However, inter-rater agreement is much higher when we perform the calculation only for contributors with a high reliability: the value of alpha is 0.40 (resp. 0.76) when we consider contributors with R greater than 0.5 (resp. 0.7).

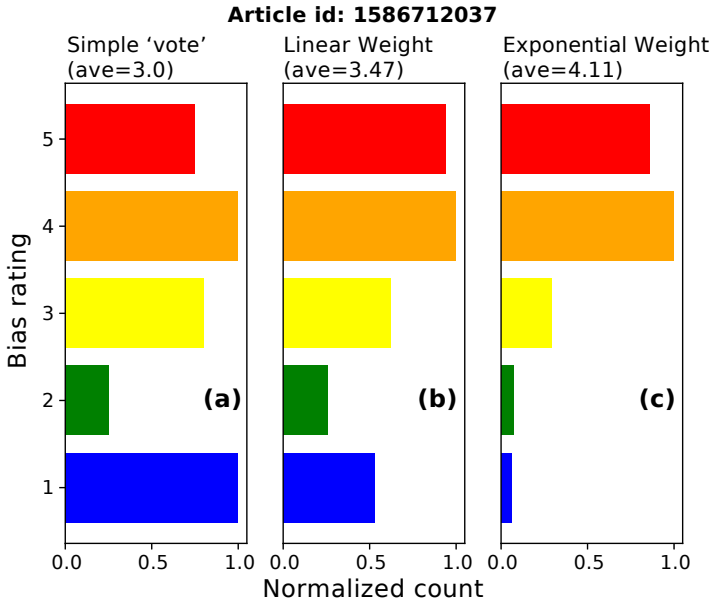


Fig. 2. Histogram displaying the bias ratings collected for an article titled "The invasion of Canada" (a) simple count of the number of users who provided each rating, (b) count weighted by users' reliability and (c) count exponentially weighted by users' reliability as explained in the text.

4.2 Assessing articles' bias based on contributors' ratings

Our goal is to determine the articles' bias and a degree of confidence in that classification based on signals provided by the crowd. A straightforward way to obtain an overall rating is to simply take each assessment as a 'vote' and average these to obtain a single value for the article.

However to try and get closer to an objective assessment of the article's bias, we tested the approach of weighting each rating by the reliability of the contributor. We tested a 'linear' weight for which a user's rating is weighted by its reliability R and a more aggressive 'exponential' weight for which a user's rating is weighted by $10^{4\times(R-1/2)}$ so that an absolutely reliable ($R = 1$) contributor's rating would weight a hundred times more than a contributor of reliability $R = 0.5$.

Figure 2 compares an article's ratings obtained with these different weightings applied. While the article's bias appears disputed from a simple vote perspective (Fig.

2a) with as many contributors judging the article as 'unbiased' (1) and 'biased' (4), it appears quite clearly biased when the exponential weight is applied (Fig. 2c). This reflects the fact that contributors who deem the article biased have higher reliability. In this case, the weighting is improving the clarity of the data collected since this reference-free, anecdote-based article in "Breaking Israel news" on "the invasion of Canada" by "hordes of illegal aliens from Syria, Haiti and anywhere else" can arguably be classified as biased.

5 Experiment: using this annotated dataset to improve the machine learning model

At Factmata we have a model to detect extreme political content online, which we provide as part of our commercial offering. One of the machine learning models was trained on a corpus of 35,236 articles scraped from domains that came from an open-source list of highly biased domains. This training dataset has noisy labels, so we decided to use the new labelled dataset described in this paper to estimate the performance of our algorithm, as well as understand how the performance would change if we added this dataset to the training data.

We first quantized the aggregated weighted scores, so that each article would fall into one of three categories: "very biased"", "unbiased" or "mixed/undecided". We only kept the first two categories, so we ended up with 280 biased instances (i.e. positives) and 260 unbiased instances (i.e. negatives). We split this dataset into training and test, splitting by domains. A domain tends to use similar language across all its pages, so by creating this test set, we are measuring how well a model generalizes to a new unseen domain. We ended up with the dataset described in the table below.

Dataset	Number of positives	Number of negatives
Original training	8971	26265
Original + manual training	9133	26439
Manual test dataset	86	118

We ran an experiment, where we trained the model on the aggregated training dataset, as well as the original. We then measured the performance improvement on the manually labeled test set. The results are in the table below:

Performance metrics on manual test set	Precision	Recall	F1-score	ROC-AUC
Original training	0.74	0.78	0.76	0.52

Original + manual training	0.71	0.88	0.78	0.65

As we can see, the largest improvement was seen in the recall of the new model, likely because the manually labeled dataset has captured types of political bias that do not occur in the open-source dataset. Even though we only increased the training data by less than 1%, the ROC-AUC improved by 25% and the recall by 13%. This is a promising result showing that a small addition of manually labeled data can make a significant improvement in the predictive power of a model trained on noisy labels.

6 Conclusion, and future work

In this paper, we have presented work in progress to create a corpus of news articles labeled for political bias and development of a method to identify reliable contributors. As a first step, we compute a reliability score for each contributor by comparing their assessment to a set of experts-created acceptable assessments on a subset of the articles. Using a probabilistic framework allows us to estimate the confidence we can have in users' reliability scores. Weighting users' contributions by their reliability score increases the clarity of the data and allows us to identify the articles that have been confidently classified by the consensus of high reliability users to train our machine learning algorithms. This notably allows us to note that high reliability contributors disagree on the bias rating for about a third of the articles, which we use to train our machine learning model to recognize uncategorizable articles in addition to biased and unbiased.

This research is very preliminary. An important next step will be to learn about potential contributors' bias from the pattern of their article ratings: for instance a contributor might be systematically providing more "left-leaning" or "right-leaning" ratings than others, which could be taken into account as an additional way to generate objective classifications. This would turn a low quality input into useful data. Another avenue of research will be to mitigate possible bias in the gold dataset. This can be achieved by broadening the set of experts providing acceptable classification and/or by also calculating a reliability score for experts, who would start with a high prior reliability but have their reliability decrease if their ratings diverge from a classification by other users when a consensus emerges.

References

1. Budak, C., Goel, S. and Rao, J.M., 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), pp.250-271.
2. Ismail, R. and Josang, A., 2002. The beta reputation system. *BLED 2002 proceedings*, p.41.
3. Krippendorff, K., 2011. Computing Krippendorff's alpha-reliability.
4. Lazaridou, K. and Krestel, R., 2016. Identifying Political Bias in News Articles. *Bulletin of the IEEE TCDL*, *12*.

5. Martin, G.J. and Yurukoglu, A., 2017. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9), pp.2565-99.
6. Marwick, A. and Lewis, R., 2017. Media manipulation and disinformation online. New York: Data & Society Research Institute.
7. Minar, M.R. and Naher, J., 2018. Violence originated from Facebook: A case study in Bangladesh. arXiv preprint arXiv:1804.11241.
8. Patankar, A.A. and Bose, J., 2016, June. Bias Based Navigation for News Articles and Media. In *International Conference on Applications of Natural Language to Information Systems* (pp. 465-470). Springer, Cham.
9. Swamynathan, G., Almeroth, K.C. and Zhao, B.Y., 2010. The design of a reliable reputation system. *Electronic Commerce Research*, *10*(3-4), pp.239-270.

Appendix

1. Article bias assessment instructions provided to contributors

Definition

Biased articles provide an unbalanced point of view in describing events; they are either strongly opposed to or strongly in favor of a person, a party, a country… Very often the bias is about politics (e.g. the article is strongly biased in favor of Republicans or Democrats), but it can be about other entities (e.g. anti-science bias, pro-Brexit bias, bias against a country, a religion…).

A biased article supports a particular position, political view, person or organization with overly suggestive support or opposition with disregard for accuracy, often omitting valid information that would run counter to its narrative.

Often, extremely biased articles attempt to inflame emotion using loaded language and offensive words to target and belittle the people, institutions, or political affiliations it dislikes.

Rules and Tips

Rate the article on the "bias scale" following these instructions:

- Provide a **rating of 1** if the article is **not biased at all**; the article might discuss cooking, movies, lifestyle… or talk about politics in a neutral and factual way.
- Provide a **rating of 2** if the article is **fairly unbiased**; the article might talk about contentious topics, like politics, but remains fairly neutral.
- Provide a **rating of 3** if the article is **somewhat biased** or if it is impossible to determine its bias, or the article is ambivalent (i.e. biased both for and against the same entity).
- Provide a **rating of 4** if the article is **clearly biased**; it overtly favors or denigrates a side, typically an opinion piece with little fairness.
- Provide a **rating of 5** if the article is **extremely biased / hyper partisan**; it overtly favors a side in emphatic terms and/or belittles the other 'side', with disregard for accuracy, and attempts to incite an action or emotion in the reader.

Please **do not include your own personal political opinion** on the subject of the article or the website itself. If you agree with the bias of the article, you still should

tag is as biased. Try and remove any sense of your personal political beliefs, and critically examine the language and the way the article has been written.
Please do not pay attention to other information on the webpage (page layout, other articles, advertising etc.). **Only the content of the article is relevant** here: text, hyperlinks in it, photos and videos within the text of the article. Also, do not look at the title of the website, its name, or how it looks - just examine the article in front of you and its text.
Do not answer randomly, we will reject submissions if there is evidence that a worker is providing spam responses. Do not skip the rating, providing an overall bias is required.

Examples

- Example of sentences from an hyper-partisan article with many mentions about Donald Trump, clear opposition towards him and loaded language in bold (such an article should be rated as 5):

"*This is how a trickle-down of **vileness** acquires a fire hose. But the big story doesn't stop with Trump's globe-wide gift to the **worst devils** of human nature. The big story is that Trump, or his trusted Ministers of Internet Intake, inhabits a **bottom-barrel** world in which Fox News and Infowars and Gateway Pundit and—sure—Britain First loom large. They're picking this stuff up, combining through it, repurposing it all the time*"

- Example of another hyper-partisan article, with loaded anti-Clinton language in bold, and a call to action at the end for others to join and support the ideology:

"It's a neat little magic trick. It is also **incredibly unethical** and most likely illegal… but then again, that never stopped **the Clinton machine** before. Please press "Share on Facebook" if you think these **dirty tricks** need to be exposed!"

- Example of a biased article (should be rated as 4 on the 1-5 scale). Here, there is less loaded language, but clearly the article is one sided towards Trump:

"*President Trump's stock market rally is historical! No President has seen more all time highs (63) in their first year in office than President Trump. President Trump set the record earlier this year for the most all time closing stock market highs during his first year in office. Currently the Dow has set 80 closing highs since last year's election and 63 since President Trump's inauguration. (As a comparison, President Obama had no stock market highs his entire first term.)*"

- Example of an article talking about a trivial topic. Even though the article speaks positively about money orders and Rite Aid, this shouldn't be marked as biased (should be rated as 1):

"*For people who want to pay bills, purchase goods, or simply want to send guaranteed funds without the risk associated with exchanging cash, money orders are a trusted method of payment. Rite Aid money orders are convenient because of the low fees, numerous locations, and long hours.*"

2. Sample annotation data

pageurl	worker_id	article_bias
url1	44278209	3.0
url1	43718845	4.0
url1	38202325	4.0
url1	37881503	4.0
url1	44164300	4.0
url1	55128002	4.0
url1	55128001	3.0
url1	55128003	2.0
url2	31613324	3.0
url2	44128742	2.0
url2	39793872	5.0
url2	38202325	5.0
url2	44303394	5.0
url2	37881503	4.0
url2	55128002	4.0
url2	55128003	4.0
url2	55128004	5.0
url3	31613324	4.0
url3	44128742	5.0
url3	16718271	1.0
url3	43951421	1.0
url3	44303394	3.0
url3	38202325	4.0
url3	37881503	2.0
url3	55128002	1.0
url3	55128001	1.0
url3	55128003	1.0
url3	55128004	1.0

- url1: http://www.stlamerican.com/news/local_news/privilege-at-the-protest-white-allies-demonstrate-without-incident-outside/article_543f4ba2-9f5f-11e7-95d0-c3a75bed0e90.html
- url2: http://www.theamericanconservative.com/buchanan/trump-embraces-the-culture-war/
- url3: http://www.thegatewaypundit.com/2017/10/breaking-active-shooter-reported-usc-campus-lockdown-videos/

CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement

Anca Dumitrache*[1], Oana Inel*[1], Lora Aroyo[1], Benjamin Timmermans[2], and Chris Welty[3]

[1] Vrije Universiteit Amsterdam
{anca.dmtrch,oana.inel,l.m.aroyo}@gmail.com
[2] CAS IBM Nederland
b.timmermans@nl.ibm.com
[3] Google AI
cawelty@gmail.com

Abstract. Typically crowdsourcing-based approaches to gather annotated data use inter-annotator agreement as a measure of quality. However, in many domains, there is ambiguity in the data, as well as a multitude of perspectives of the information examples. In this paper, we present ongoing work into the CrowdTruth metrics, that capture and interpret inter-annotator disagreement in crowdsourcing. The CrowdTruth metrics model the inter-dependency between the three main components of a crowdsourcing system – worker, input data, and annotation. The goal of the metrics is to capture the degree of ambiguity in each of these three components. The metrics are available online at https://github.com/CrowdTruth/CrowdTruth-core.

1 Introduction

The process of gathering ground truth data through human annotation is a major bottleneck in the use of information extraction methods. Crowdsourcing-based approaches are gaining popularity in the attempt to solve the issues related to volume of data and lack of annotators. Typically these practices use inter-annotator agreement as a measure of quality. However, this assumption often creates issues in practice. Previous experiments we performed [2] found that inter-annotator disagreement is usually never captured, either because the number of annotators is too small to capture the full diversity of opinion, or because the crowd data is aggregated with metrics that enforce consensus, such as majority vote. These practices create artificial data that is neither general nor reflects the ambiguity inherent in the data.

To address these issues, we proposed the **CrowdTruth** [3] method for crowdsourcing ground truth by harnessing inter-annotator disagreement. We present an alternative approach for crowdsourcing ground truth data that, instead of enforcing agreement between annotators, captures the ambiguity inherent in

* Equal contribution, authors listed alphabetically.

semantic annotation through the use of disagreement-aware metrics for aggregating crowdsourcing responses. In this paper, we introduce the second version of **CrowdTruth metrics** – a set of metrics that capture and interpret inter-annotator disagreement in crowdsourcing annotation tasks. As opposed to the first version of the metrics, published in [6], the current version models the *inter-dependency between the three main components of a crowdsourcing system – worker, input data, and annotation.* This update is based on the intuition that disagreement caused by low quality workers should not be interpreted as the data being ambiguous, but also that ambiguous input data should not be interpreted as due to the low quality of the workers.

This paper presents the definitions of the CrowdTruth metrics 2.0, together with the theoretical motivations of the updates based on the previous version 1.0. The code of the implementation of the metrics is available on the CrowdTruth Github.[4] The 2.0 version of the metrics has already been applied successfully to a number of use cases, e.g. semantic frame disambiguation [5], relation extraction from sentences [4], topic relevance [7]. In the future, we plan to continue the validation of the metrics through evaluation over different annotation tasks, comparing CrowdTruth approach with other disagreement-aware crowd aggregation methods.

2 CrowdTruth Methodology

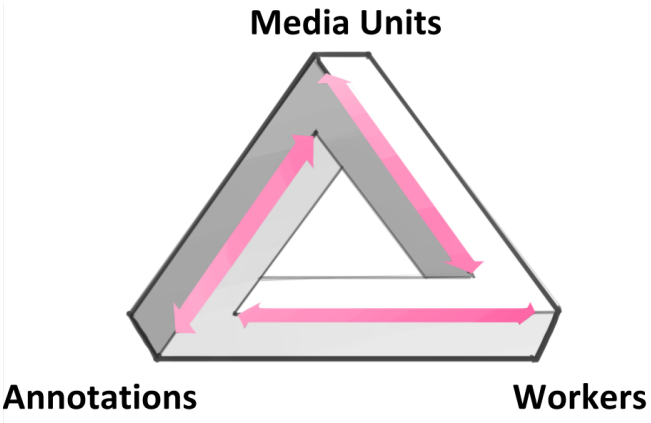


Fig. 1. Triangle of Disagreement

The CrowdTruth methodology consists of a set of quality metrics and best practices to aggregate inter-annotator agreement such that ambiguity in the task is preserved. The methodology uses the triangle of disagreement model (based on

[4] `https://github.com/CrowdTruth/CrowdTruth-core`

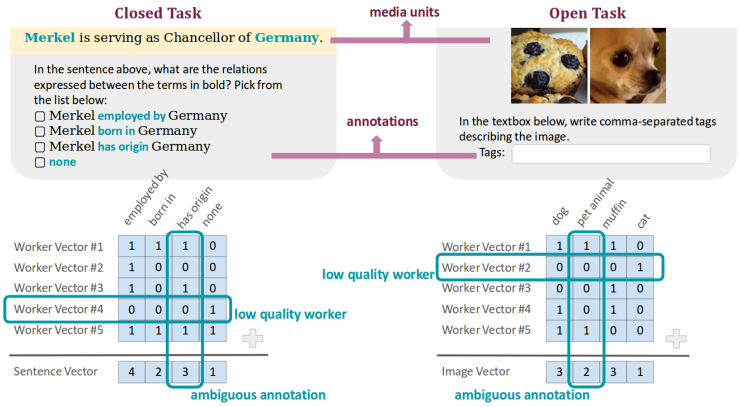


Fig. 2. Example closed and open tasks, together with the vector representations of the crowd answers.

the triangle reference [8]) to represent the crowdsourcing system and its three main components – input media units, workers, and annotations (Figure 1). The triangle model expresses how ambiguity in any of the corners disseminates and influences the other components of the triangle. For example, an unclear sentence or an ambiguous annotation scheme would cause more disagreement between workers [1], and thus, both need to be accounted for when measuring the quality of the workers.

The CrowdTruth methodology calculates quality metrics for workers, media units and annotations. The novel contribution of version 2.0 is that *the way how ambiguity propagates between the three components of the crowdsourcing system has been made explicit in the quality formulas of the components*. So for example, the quality of a worker is weighted by the quality of the media units the worker has annotated, and the quality of the annotations in the task.

This section describes the two steps of the CrowdTruth methodology:

1. formalizing the output from crowd tasks into **annotation vectors**;
2. calculating quality scores over the annotation vectors using **disagreement metrics**.

2.1 Building the Annotation Vectors

In order to measure the quality of the crowdsourced data, we need to formalize crowd annotations into a **vector space representation**. For *closed tasks*, the annotation vector contains the given answer options in the task template, which the crowd can choose from. For example, the template of a *closed task* can be composed of a multiple choice question, which appears as a list checkboxes or radio buttons, thus, having a finite list of options to choose from. Figure 2 shows an example of a closed and an open task, indicating also what the media units and annotations are for both cases.

While for *closed tasks* the number of elements in the annotation vector is known in advance, for *open-ended tasks* the number of elements in the annotation vector can only be determined when all the judgments for a media unit have been gathered. An example of such a task can be highlighting words or word phrases in a sentence, or as an input text field where the workers can introduce keywords. In this case the answer space is composed of all the unique keywords from all the workers that solved that media unit. As a consequence, all the media units in a closed task have the same answers space, while for open-ended tasks the answer space is different across all the media units. Although the answer space for open-ended tasks is not known from the beginning, it still can be further processed in a finite answer space.

In the annotation vector, each answer option is a boolean value, showing whether the worker annotated that answer or not. This allows the annotations of each worker on a given media unit to be aggregated, resulting in a **media unit vector** that represents for each option how often it was annotated. Figure 2 shows how the worker and media unit vectors are formed for both a closed and an open task.

2.2 Disagreement Metrics

Using the vector representations, we calculate three core metrics that capture the **media unit quality**, **worker quality** and **annotation quality**. These metrics are mutually dependent (e.g. the media unit quality is weighted by the annotation quality and worker quality), based on the idea from the triangle of disagreement that ambiguity in any of the corners disseminates and influences the other components of the triangle. The mutual dependence requires an iterative dynamic programming approach, calculating the metrics in a loop until convergence is reached. All the metrics have scores in the $[0, 1]$ interval, with 0 meaning low quality and 1 meaning high quality. Before starting the iterative dynamic programming approach, the quality metrics are initialized with 1.

To define the CrowdTruth metrics, we introduce the following notation:

- $workers(u)$: all workers that annotate media unit u;
- $units(i)$: all input media units annotated by worker i;
- $WorkVec(i, u)$: annotations of worker i on media unit u as a binary vector;
- $MediaUnitVec(s) = \sum_{i \in workers(s)} WorkVec(i, s)$, where s is an input media unit.

To calculate agreement between 2 workers on the same media unit, we compute the cosine similarity over the 2 worker vectors. In order to reflect the dependency of the agreement on the degree of clarity of the annotations, we compute $Wcos$, the weighted version of the cosine similarity. The Annotation Quality Score (AQS), which will be described in more detail at the end of the section, is used as the weight. For open-ended tasks, where annotation quality cannot be calculated across multiple media units, we consider annotation quality equal to 1 (the maximum value) in all cases. Given 2 worker vectors, vec_1 and vec_2 on the same media unit, the formula for the weighted cosine score is:

$$Wcos(vec_1, vec_2) =$$
$$= \frac{\sum_a vec_1(a)\ vec_2(a)\ AQS(a)}{\sqrt{(\sum_a vec_1^2(a)\ AQS(a))\ (\sum_a vec_2^2(a)\ AQS(a))}},$$
$$\forall a \text{ - annotation.}$$

The **Media Unit Quality Score (UQS)** expresses the overall worker agreement over one media unit. Given an input media unit u, $UQS(u)$ is computed as the average cosine similarity between all worker vectors, weighted by the worker quality (WQS) and annotation quality (AQS). Through the weighted average, workers and annotations with lower quality will have less of an impact on the final score. The formula used in its calculation is:

$$UQS(u) = \frac{\sum_{i,j} WorkVecWcos(i,j,u)\ WQS(i)\ WQS(j)}{\sum_{i,j} WQS(i)\ WQS(j)},$$

$$WorkVecWcos(i,j,u) = Wcos(WorkVec(i,u), WorkVec(j,u)),$$
$$\forall i,j \in workers(u), i \neq j.$$

The **Worker Quality Score (WQS)** measures the overall agreement of one crowd worker with the other workers. Given a worker i, $WQS(i)$ is the product of 2 separate metrics - the worker-worker agreement $WWA(i)$ and the worker-media unit agreement $WUA(i)$:

$$WQS(i) = WUA(i)\ WWA(i).$$

The **Worker-Worker Agreement (WWA)** for a given worker i measures the average pairwise agreement between i and all other workers, across all media units they annotated in common, indicating how close a worker performs compared to workers solving the same task. The metric gives an indication as to whether there are consistently like-minded workers. This is useful for identifying communities of thought. $WWA(i)$ is the average cosine distance between the annotations of a worker i and all other workers that have worked on the same media units as worker i, weighted by the worker and annotation qualities. Through the weighted average, workers and annotations with lower quality will have less of an impact on the final score of the given worker.

$$WWA(i) =$$
$$\frac{\sum_{j,u} WorkVecWcos(i,j,u)\ WQS(j)\ UQS(u)}{\sum_{j,u} WQS(j)\ UQS(u)},$$
$$\forall j \in workers(u \in units(i)), i \neq j.$$

The **Worker-Media Unit Agreement (WUA)** measures the similarity between the annotations of a worker and the aggregated annotations of the rest of the workers. In contrast to the WWA which calculates agreement with individual workers, WUA calculates the agreement with the consensus over all workers. $WUA(i)$ is the average cosine distance between the annotations of a worker i and all annotations for the media units they have worked on, weighted by the media unit (UQS) and annotation quality (AQS). Through the weighted average, media units and annotations with lower quality will have less of an impact on the final score.

$$WUA(i) = \frac{\sum_{u \in units(i)} WorkUnitWcos(u,i)\ UQS(u)}{\sum_{u \in units(i)} UQS(u)},$$

$$WorkUnitWcos(u,i) = Wcos(WorkVec(i,u), \\ MediaUnitVec(u) - WorkVec(i,u))$$

The **Annotation Quality Score (AQS)** measures the agreement over an annotation in all media units that it appears. Therefore, it is only applicable to closed tasks, where the same annotation set is used for all input media units. It is based on $P_a(i|j)$, the probability that if a worker j annotates a in a media unit, worker i will also annotate it.

$$P_a(i|j) = \frac{\sum_u UQS(u)\ WorkVec(i,s)[a]\ WorkVec(j,s)[a]}{\sum_u UQS(u)\ WorkVec(j,u)(r)},$$
$$\forall u \in units(i) \cap units(j).$$

Given an annotation a, $AQS(a)$ is the weighted average of $P_a(i|j)$ for all possible pairs of workers i and j. Through the weighted average, input media units and workers with lower quality will have less of an impact on the final score of the annotation.

$$AQS(a) = \frac{\sum_{i,j} WQS(i)\ WQS(j)\ P_a(i|j)}{\sum_{i,j} WQS(i)\ WQS(j)},$$
$$\forall i,j \text{ workers}, i \neq j.$$

The formulas for media unit, worker and annotation quality are all mutually dependent. To calculate them, we apply an iterative dynamic programming approach. First, we initialize each quality metric with the score for maximum quality (i.e. equal to 1). Then we repeatedly re-calculate the quality metrics until each of the values are stabilized. This is assessed by calculating the sum of

variations between iterations for all quality values, and checking until it drops under a set threshold t.

The final metric we calculate is the **Media Unit - Annotation Score (UAS)** – the degree of clarity with which an annotation is expressed in a unit. Given an annotation a and a media unit u, $UAS(u,a)$ is the ratio of the number of workers that picked annotation u over all workers that annotated the unit, weighted by the worker quality.

$$UAS(u,a) = \frac{\sum_{i \in workers(u)} WorkVec(i,u)(a)\ WQS(i)}{\sum_{i \in workers(u)} WQS(i)}.$$

3 Conclusion

In this paper, we present ongoing work into the CrowdTruth metrics, that capture and interpret inter-annotator disagreement in crowdsourcing. Typically crowdsourcing-based approaches to gather annotated data use inter-annotator agreement as a measure of quality. However, in many domains, there is ambiguity in the data, as well as a multitude of perspectives of the information examples. The CrowdTruth metrics model the inter-dependency between the three main components of a crowdsourcing system – worker, input data, and annotation.

We have presented the definitions and formulas of several CrowdTruth metrics, including the three core metrics measuring the quality of workers, annotations, and input media units. The metrics are based on the idea of the triangle of disagreement, expressing how ambiguity in any of the corners disseminates and influences the other components of the triangle. Because of this, disagreement caused by low quality workers should not be interpreted as the data being ambiguous, but also that ambiguous input data should not be interpreted as due to the low quality of the workers. The metrics have already been applied successfully to use cases in topic relevance [7], semantic frame disambiguation [5] and relation extraction from sentences [4].

References

1. Aroyo, L., Welty, C.: The Three Sides of CrowdTruth. Journal of Human Computation **1**, 31–34 (2014)
2. Aroyo, L., Welty, C.: Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. Web Science 2013. ACM (2013)
3. Aroyo, L., Welty, C.: Truth Is a Lie: CrowdTruth and the Seven Myths of Human Annotation. AI Magazine **36**(1), 15–24 (2015)
4. Dumitrache, A., Aroyo, L., Welty, C.: False positive and cross-relation signals in distant supervision data (2017)
5. Dumitrache, A., Aroyo, L., Welty, C.: Capturing ambiguity in crowdsourcing frame disambiguation (2018)

6. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., Sips, R.J.: Crowdtruth: Machine-human computation framework for sing disagreement in gathering annotated data. In: The Semantic Web–ISWC 2014, pp. 486–504. Springer (2014)
7. Inel, O., Li, D., Haralabopoulos, G., Van Gysel, C., Szlvik, Z., Simperl, E., Kanoulas, E., Aroyo, L.: Studying topical relevance with evidence-based crowdsourcing. In: To Appear in the Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM (2018)
8. Knowlton, J.Q.: On the definition of "picture". AV Communication Review **14**(2), 157–183 (1966)

A Case for a Range of Acceptable Annotations

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng

Google, 1600 Amphitheatre Parkway, Mountain View, California 94043
{jpalomaki, orhinehart, michaeltseng}@google.com

Abstract. Multi-way annotation is often used to ensure data quality in crowdsourced annotation tasks. Each item is annotated redundantly and the contributors' judgments are converted into a single "ground truth" label or more complex annotation through a resolution technique (e.g., on the basis of majority or plurality). Recent crowdsourcing research has argued against the notion of a single "ground truth" annotation for items in semantically oriented tasks—that is, we should accept the aggregated judgments of a large pool of crowd contributors as "crowd truth." While we agree that many semantically oriented tasks are inherently subjective, we do not go so far as to trust the judgments of the crowd in all cases. We recognize that there may be items for which there is truly only one acceptable response, and that there may be divergent annotations that are truly of unacceptable quality. We propose that there exists a class of annotations between these two categories that exhibit *acceptable variation*, which we define as the range of annotations for a given item that meet the standard of quality for a task. We illustrate acceptable variation within existing annotated data sets, including a labeled sound corpus and a medical relation extraction corpus. Finally, we explore the implications of acceptable variation on annotation task design and annotation quality evaluation.

Keywords: crowdsourcing · human computation · ambiguity · disagreement.

1 Introduction

With respect to annotation quality, one dichotomy in human-annotated data is the categorization of crowd contributors' annotations or labels into "noise" and the more nebulous notion of "ground truth." The development of a human-annotated "gold" standard has long been seen as essential to the training and evaluation of natural language processing systems in particular.

Recent research suggests that there is no such thing as a single "golden" label or a single ground truth for semantically oriented data [3, 4, 2]. Factors such as the ambiguity of the input items and clarity of task guidelines, as well as differences in contributor backgrounds and levels of conservativeness, have been shown to condition disagreement among annotations [6, 8, 9].

With respect to noise, whether annotations are provided by "expert" annotators, crowd contributors, or some other source, "spammy" or noisy data is a pervasive and perhaps inevitable feature of human-annotated data.

Regarding ground truth, it may be that for some subset of items paired with a given semantic annotation task, there really is only one label that satisfies the standard of quality for the task. This would likely be the case if for that subset of input items, the task guidelines are clear, the content of the items themselves is less ambiguous, and contributor-specific factors such as background would not condition varying interpretations of the task or items.

So then what lies between noise and the notion of a single "golden" label for each item in a data set for a given task? We propose here that this middle ground is occupied by annotations that exhibit what we call acceptable variation. We define acceptable variation as the range of annotations (labels, answers, etc.) at any annotation stage that meet the standard of quality for a given task, which we conceptualize in terms put forth by Allahbakhsh et al. [1], who define quality for tasks as "the extent to which the provided outcome fulfills the requirements of the requester."

We will illustrate, using examples from previous semantic annotation tasks that have been used by researchers to argue against the notion of a single ground truth, that there is a kind of disagreement which falls into the range of acceptable variation. We will extend our argument to non-semantic input data to show that the notion of acceptable variation is useful to any annotation task requiring human judgments.

Finally, we will discuss the implications of our proposal. If we accept that for a given annotation task, there is some subset of input items that allow a range of acceptably varying annotations, then we must be explicit about how this affects our orientation to task design, the evaluation of the annotations, and thus the overall quality of the labeled data. We argue that acceptable variation is an important feature of any data set that is meant to be representative of the kind of complex semantic phenomena that require human annotation. We propose that task design and task guidelines should be broad enough to allow for a wealth of answers and thereby facilitate gathering more representative data. We will also discuss ways in which acceptable variation can be identified and leveraged.

Furthermore, we will argue that the existence of acceptable variation has important implications for how we evaluate the performance of contributors. Task designers must be careful not to mistake acceptable variation for noise and unwittingly penalize contributors who are providing us with valuable signals about the complexity of the phenomena we are trying to understand.

2 Conditioning Acceptable Variation

It is worth considering what gives rise to both disagreement and the particular type of disagreement that we characterize as acceptable variation. We follow Dumitrache [6] and Kairam and Heer [8] who discuss three main sources of disagreement: (1) differences in contributors, including background knowledge or understanding of the task, which may lead them to annotate more or less

conservatively; (2) clarity of the expected annotation label(s); and (3) ambiguity of the content of the task's input items from the data set.

We will show that these factors are important in understanding how acceptable variation is conditioned, and we will demonstrate examples of each type in the annotation tasks we consider below.

3 Acceptable Variation in Published Data Sets

Aroyo and Welty [3] introduce "crowd truth" as an alternative to ground truth (the notion that for each task item there is such as a thing as a single correct annotation). They ground their discussion of crowd truth in the disparity in annotator disagreement between entity type annotation tasks and relation annotation tasks. They note that disagreement tends to be much higher for relation annotation tasks because there are multiple ways to express the same relation and, conversely, the same linguistic expression may express many different relations.

Aroyo and Welty [3, 4] argue that the disagreement seen among contributors in relation annotation tasks is not necessarily noise or a sign of a poorly defined problem, but rather a useful and informative property of human-annotated data that provides a signal about the vagueness or ambiguity of the input items for the annotation task.

In this section, we will adduce examples of acceptable variation in a medical relation corpus and a labeled sound corpus, which are multi-classification and free response tasks, respectively. We acknowledge that acceptable variation may surface in other types of tasks as well, such as ones with binary or scale rating (e.g., sentiment analysis). We may explore those in later work.

3.1 Medical Relation Corpus

The medical relation annotation task that Aroyo and Welty [4] used to investigate the disagreement in medical relation extraction allowed crowd contributors to select from a given set any relations that they judge as applicable to highlighted terms in a given sentence. An additional step asked contributors to provide a justification for their selection of the relation(s) they chose for each input sentence. The set of relations was selected manually from the United Medical Languages System (UMLS). The set included the relations *treats*, *prevents*, *diagnosed by test or drug*, *causes*, *location*, *symptom*, *manifestation*, *contraindicates*, *associated with*, *side effect*, *is a*, *part of*, *other*, and *none*. Crowd contributors were provided with a definition and example sentence for each relation (and given the design of the task, it is likely that the terms were interpreted in their local linguistic contexts). Each input sentence was annotated by multiple contributors.

The annotations demonstrate the kind of disagreement we characterize as acceptable variation, which we differentiate from noise or spammy annotations.

Consider the example task item in Table 1 with original highlighted terms in capital letters. The task item elicits varying annotations, both within and outside of the acceptable range. Note that since contributors were allowed to select all relations that they judged to be applicable, the number of judgments (16) exceeds the number of contributors (15) who annotated this sentence.

Table 1. Counts of labels chosen by crowd contributors for the relations between "RIBOFLAVIN" and "RIBOFLAVIN DEFICIENCY" for this given sentence from the medical relation annotation task.

These data suggest that subclinical RIBOFLAVIN DEFICIENCY may occur in adolescents and that deficiency may be related to dietary intake of RIBOFLAVIN.	
Relation Annotation	**Count**
associated with	4
symptom	3
causes	3
prevents	1
side effect	1
manifestation	1
part of	1
diagnose by test or drug	1
other	1

Not all of the relation annotations for the example above can be characterized as falling within the acceptable range. For example, it is unlikely that (the dietary intake of) "RIBOFLAVIN" is a *manifestation* of "RIBOFLAVIN DEFICIENCY." Yet the under-specification of what kind of "dietary intake" of "RIBOFLAVIN" is related to "RIBOFLAVIN DEFICIENCY" opens the sentence up to different interpretations. The sentence is unclear as to whether "dietary intake of riboflavin" here refers to lack of, insufficient, sufficient, or excessive consumption of riboflavin. While "riboflavin deficiency" can suggest that the relation is conditioned by lack of (or insufficient) riboflavin consumption, the variation of annotations suggests that this condition is not obvious to crowd contributors who are not medical experts. This is unsurprising given that contributor backgrounds have been shown by previous research [6, 8] to condition disagreement. It is worth noting that Aroyo and Welty [3, 4] found that even medical experts had difficulty reaching consensus on what relations held between terms in similar sentences.

Let us assume that riboflavin deficiency is caused by the lack of (or insufficient) consumption of riboflavin. Considering the other factors that Dumitrache [6] and Kairam and Heer [8] point to as conditioning disagreement, the relation annotation *associated with* may have been selected by more conservative annotators, as it subsumes a relationship in which any degree of riboflavin consumption may be causally related to riboflavin deficiency. The annotation falls within the range of acceptable annotations for this task.

If we interpret the phrase "dietary intake of riboflavin" as referring to the medically recommended level of riboflavin consumption, then the relation annotation *prevents* also falls within the range of acceptable annotations for this task. The contributor who selected this relation may have had prior knowledge about the relationship between riboflavin and riboflavin deficiency, or may have been more liberal in the interpretation of the relationship suggested by the sentence.

If we interpret "dietary intake of riboflavin" as referring to the lack of (or insufficient) consumption of riboflavin, we can understand why the *causes* and *symptom* and even *side effect* relations were selected so frequently, regardless of the fact that the relation between the individual highlighted terms is not best expressed by these labels. The relations *causes*, *symptom*, and *side effect* may not hold between the specific terms "RIBOFLAVIN" and "RIBOFLAVIN DEFICIENCY," yet a causal relation is indeed suggested by the sentence. These annotations may fall outside of the range of acceptable variation, but they provide a valuable signal of how the input sentence and the terms themselves may lead to ambiguous interpretations by crowd contributors.

The frequency with which contributors selected the *causes* and *symptom* relation annotations places them within the set of relation annotations selected by a plurality of contributors. These judgments are therefore in some sense "true" under the notion of crowd truth, as they depend upon (and provide insight into) the ambiguity of the input sentences as interpreted by the crowd contributors for this task. Acceptable variation is distinct from crowd truth in that it is independent of plurality or majority. For a given input item for a multi-way annotation task, it may be the case that only one contributor selects a label or annotation, yet this choice may fall within the range of acceptable variation, as with the *prevents* relation annotation for the input sentence in Table 1.

Even seemingly straightforward input items can condition acceptable variation. Consider the example in Table 2.

Since the input sentence contains the relation expression "caused," it is not surprising that the majority of contributors selected the relation annotation *causes*. However, the broader relation annotation *associated with* subsumes the *causes* relation, so it also falls within the range of acceptable variation for this task. This is notable because this relation annotation was selected by only one contributor out of 15, suggesting that at least for this particular input item, majority and plurality are not necessarily indicators of the quality or "truth" of annotations.

Table 2. Counts of labels chosen by crowd contributors for the relations between "FUNGI" and "FUNGAL INFECTIONS" for this given sentence from the medical relation annotation task.

FUNGAL INFECTIONS may be caused by several FUNGI the most important of these being Candida species including C. albicans, C. glabrata, C. krusei, C. tropicalis, C. parapsilosis, and C. guillermondii.	
Relation Annotation	**Count**
causes	13
associated with	1
part of	2

3.2 VU Sound Corpus

Crowdsourced annotation tasks can be designed to encourage a variety of acceptable answers. An example of such is the labeling task that generated the VU Sound Corpus [12]. In this task, contributors added keyword descriptions to environmental sound clips from the Freesound database [7]. The authors used quality analysis to root out spammy responses, but otherwise did not consider there to be "right" or "wrong" labels for any given sound clip. Rather than restrict the format of the labels in the task's guidelines or user interface, post-processing was used to normalize and cluster keywords, which helped identify annotation outliers.

We will show that acceptable variation is present in the VU Sound Corpus, and that it is likely conditioned by factors previously discussed, including differences in contributor background and conservativeness. Furthermore, we will discuss how the presence of acceptable variation is a desirable attribute of the labeled data.

Table 3 is an example where contributors' background knowledge may have conditioned the variation seen in the responses. Some contributors (8, 10) added keywords referencing a possible source of the sound, *feedback*, while others (6, 7) opted for more concrete keywords describing the sound itself, such as *whistle*. This difference in emphasis may be due to varying levels of familiarity or world experience with how to generate a sound of this type ("eerie horror film sound").

Table 4 demonstrates differing levels of conservativeness of contributors. The input sound clip of a hand-held electric beater in use is ambiguous enough that contributors who chose to add specific labels for motorized devices split into two groups: those who used labels related to *drill* (2, 3, 5, 7, 8, 9), and those who used labels related to *blender* (4, 6; 9 falls into both groups if *mixer* is understood as a synonym of *blender*). Two contributors (1, 10) opted to back off to a more general keyword, *machine*. Van Miltenberg et al. acknowledge that, while keywords related to *drill* and *blender* both incorrectly identify the source

Table 3. Lists of keywords applied by each crowd contributor (identified by number) to the given "eerie horror film sound" audio clip from the VU Sound Corpus labeling task.

URL	https://www.freesound.org/people/NoiseCollector/sounds/6212/
Description	Multisamples created with subsynth. Eerie horror film sound in middle and higher registers. Normalized and converted to AIFF in cool edit 96. File name indicates frequency for example: HORROC04.aif= C4, where last 3 characters are C04
Tags	evil, horror, subtractive, synthesis
Keywords	
1	swing, metallic
2	shrill, shriek
3	whine
4	high, tuning, resonance
5	HIGH PITCHED SOUND
6	whistle
7	wistle
8	amplifier feedback, high pitched tone
9	screech
10	feedback

of the sound, the labels still offer useful information about what the recording sounds like, which could be useful for grounded semantic models. Note that the factually incorrect label *drill* was chosen by a majority of crowd contributors: this is an example of "crowd truth" that may indicate what many in the crowd actually perceive when they hear the audio clip. The more conservative label *machine* would fall into the range of acceptable variation.

The crowd-generated keywords complement the ones generated by the authors who originally uploaded the sounds by providing perspective from contributors who were not involved with the creation of the underlying corpus of sound clips. In reference to their own labels, the authors note, "Well-informed parties commonly overlook things that are obvious to them"—that is, the designers of the corpus suffer from a "curse of knowledge" that constrains the keyword options that they would consider for an ambiguous sound.

Table 4. pubLists of keywords applied by each crowd contributor (identified by number) to the given "Sunbeam Beater-mix Pro 320 Watt electric beater" audio clip from the VU Sound Corpus labeling task.

URL	https://www.freesound.org/people/terminal/sounds/22795/
Description	A sample of my Sunbeam Beater-mix Pro 320 Watt electric beater at low speed setting #2. Recorded on 2 tracks in The Closet using a Rode NT1A and a Rode NT3 mic, mixed to stereo and processed through a multi band limiter.
Tags	appliance, beater, electric, kitchen
Keywords	
1	machine
2	drilling, grating, noisy
3	DRILL
4	blender
5	drill
6	blender
7	drill, drilling
8	drill, rattle, buzz
9	mixer,drill,whirring
10	machine

4 Task Design Implications

We can extrapolate the application of this insight to anyone who is designing an annotation task. Even if they do not participate in the creation of the underlying data set to be labeled, task designers usually have knowledge and contextual information about the data, the problem space, and the model informing the annotation scheme. If task designers constrain the task definition (guidelines, user interface, annotation options, etc.) and do not provide space for contributors to include new or unexpected annotations, then they risk foregoing the potential insights of the crowd. That is, the crowd contributors' lack of context enables them to serve as a fresh set of eyes (or ears) when evaluating the task input items and making annotation judgments. In addition, task designers should consider including channels for feedback from contributors on the task design and input

data, which may take the form of a comment box for text input within the task itself. Especially if leveraged in the earlier stages of annotation, this method can help in shaping the ultimate design of a task.

We recognize that not all task designers would be comfortable with a fully open-ended label set. As a compromise, task designers could implement an iterative process to better define the range of acceptable variation: first, run more open-ended pilots to get a sense of the range of annotations that contributors provide for a given task; next, classify the resulting annotations that diverge from expected options as acceptable or not; and then refine the annotation scheme to include new options corresponding to the acceptably varying annotations.

This approach is similar to the transition along the continuum from "user-driven" to "model-driven" annotation tasks described in Chang et al. [5] in exploring the domain of conceptual relationships implicitly expressed by noun–noun compounds. The first version of task consisted of an explicitly user-driven design encouraging contributors to write their own paraphrases describing the relationship(s) between the nouns in each given noun–noun compound. After reviewing the variety of paraphrases, the task designers settled on a more consistent format for writing the paraphrases that would allow for contributor creativity while also facilitating more reliable extraction of relationship terms that could be included in training data for a machine learning model. The task designers additionally developed more extensive guidelines and reinforcement training for contributors.

Involving crowd contributors in the iterative design of the task can also increase their understanding of the data and foster a sense of shared purpose in the data labeling effort. We encourage the research community to prioritize such engagement as an explicit component of the task design. For example, the standard practice of injecting items with "golden answers" (that is, task items that have been labeled by the task designers or other experts) into each contributor's task queue is used to evaluate the crowd's individual and aggregate performance against expert performance, but it does not necessarily measure each contributor's conscientiousness or engagement with the task.

Apparent divergence from the expert standard would be particularly exaggerated when the "golden answers" are sampled to over-represent edge cases or otherwise tricky items (this is often done when expert judgments are expensive or difficult to acquire). Furthermore, as we have discussed, for tasks where acceptable variation is expected, there is often no single correct label or annotation for a given task item, and comparing crowd contributors with experts may unfairly penalize contributors who are providing valuable information through their varying judgments. For such tasks, we would like to advance the notion of "golden *questions*"—that is, prompts that are designed to elicit consistent answers or labels. These more objective prompts need not be related to the more subjective task; they could be interspersed into the queue as special task items or appear as a section of the main task template. They would serve to establish a baseline for each crowd contributor's consistency in judgment. Contributors who answer the more objective prompts consistently (with themselves and with oth-

ers) are likely to be providing useful signal if they exhibit variation in responses to the more subjective prompts. Analysis would still be needed to distinguish the variation due to sub-optimal design of the subjective prompts from the variation that would be considered acceptable.

5 Conclusion and Future Work

We have introduced the notion of *acceptable variation*, which we conceptualize as both a natural progression of and complementary to the notion that there is no such thing as a single ground truth or "golden" label for items in semantically oriented tasks. We exhibited the existence of acceptable variation in existing research, showing that the notion is amenable to tasks that range from more inherently semantic (relation annotation) to more inherently perceptual (sound clip labeling).

5.1 Identifying and Leveraging Acceptable Variation

If we accept the notion that for a given task that requires human annotation or evaluation, some subset of items can be expected to exhibit acceptable variation, we must determine how we will differentiate acceptable variation from actual noise and how we will extract value from this distinction. This may be particularly challenging for crowdsourcing workflows that depend upon plurality-based resolution, since, as we demonstrated with examples above, it may be the case that an acceptably varying annotation was provided by a single contributor in multi-way annotation for a given input item. However, as Aroyo and Welty [4] demonstrate, the sum of different contributor disagreement measures can be used to identify contributors providing annotations of lower quality. These annotations would likely fall outside of the range of acceptable variation.

For the medical relation annotation task described above, Aroyo and Welty found that low-quality annotations were provided by contributors who disagreed consistently with other contributors across tasks, while disagreement across annotations for an individual sentence was used to score each input item for sentence clarity. Contributor annotations were also evaluated on the basis of whether contributors provided original justifications for the relations they chose for a given input sentence (rather than just copying and pasting the sentence itself in part or in whole) and on the basis of the average number of relations a contributor selected for each sentence. The authors determined that contributors attempting to appear more agreeable would consistently select multiple relations for each sentence. These three metrics were used to classify (with 98% accuracy) 12 out of 110 contributors as providers of low-quality annotations.

Aroyo and Welty's approach to identifying contributors providing low-quality annotations suggests that, for the remaining contributors, disagreement is likely to provide signal about the ambiguity of the input sentences or relation labels or other factors that might condition disagreement, such as contributors' backgrounds. We propose that similar approaches to disagreement can also be used

to identify contributors whose disagreement is likely to fall within the range of acceptable variation.

Alternatively, agreement may also provide signal about which contributors are likely to provide annotations within the acceptable range. Injecting tasks with input items that are more likely to elicit agreement in multi-way annotation (because they are less ambiguous and less likely to condition disagreement due to worker differences) could serve to establish a baseline for determining trusted contributors whose annotations for items that are likely to elicit disagreement (because they are more ambiguous) would likely fall within the acceptable range.

This type of validation method, in which verifiable annotations are used to validate subjective annotations, was employed by Kittur, Chi, and Suh [10] in experiments used to assess the utility of the micro-task market on Amazon's Mechanical Turk platform as a way to collect user input.

We argue that acceptable variation is an important feature of human-annotated data and that it is important to distinguish disagreement caused by actual annotation errors and disagreement that falls within the acceptable range. The immediate value of making the distinction is that it facilitates a more nuanced understanding of disagreement for human annotation tasks generally. Furthermore, it prompts consideration of what acceptable variation would look like for specific tasks and how it should be accounted for in task design, annotation quality evaluation, and data quality evaluation.

5.2 Acceptable Variation and Annotation Quality

The notion that there is no such thing as a single "golden" label for items in semantically oriented tasks has implications for how we measure and report contributor reliability. If we take majority or plurality as the measure of correctness and penalize contributors whose annotations fall outside of those metrics, we may be punishing contributors for annotations that fall within the range of acceptable variation. As we demonstrated, acceptable variation is not necessarily reflected by plurality or even by a majority of annotations. This suggests that even 90% agreement for an item does not guarantee that the item is unambiguous to the degree that it will not condition some acceptable variation, at least for the medical relation annotation task we examined.

Kittur et al. [11] propose that the future of crowd work should articulate a fair vision through innovation, which addresses the challenge of creating reputation systems that are not amenable to cheating or gaming while maintaining the benefits of pseudonymity and low-transaction cost hiring. Our findings suggest that to that end, task designers (and everyone else who is using or evaluating the labeled data) must also consider how to differentiate between genuine errors and acceptable variation, particularly if and when they are statistically indistinguishable, to ensure that contributor reputations are fairly and accurately assessed.

We plan to address annotation quality evaluation in future work. In our own labeled data sets, we have been able to identify acceptably varying annotations

manually on a relatively small subset of the data, but we will need to consider how to design automatic methods at scale.

We invite the research community to embrace acceptable variation as a useful insight into the complexity of human judgment for ambiguous or otherwise subjective tasks, and to confront the implications explicitly when designing and evaluating crowdsourcing tasks.

6 Acknowledgments

We would like to thank Russell Lee-Goldman, Slav Petrov, and other colleagues in Google Research for their helpful feedback on this work as well as the reviewers for their valuable comments and suggestions. We are grateful for the contributors who annotated the data sets we analyzed.

References

1. Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., Dustdar, S.: Quality Control in Crowdsourcing Systems: Issues and Directions. IEEE Internet Computing **2**(17), 76–81 (2013)
2. Aroyo, L., Welty, C.: Truth is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AI Magazine **2**(36:1), 15–24 (2015)
3. Aroyo, L., Welty, C.: Crowd Truth: Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard. In: ACM Web Science 2013. ACM, New York, NY, USA (2013)
4. Aroyo, L., Welty, C.: Measuring Crowd Truth for Medical Relation Extraction. In: AAAI Technical Report FS-13-0, Semantics for Big Data. The AAAI Press, Palo Alto, California (2013)
5. Chang, N., Lee-Goldman, R., Tseng, M.: Linguistic Wisdom from the Crowd. In: AAAI Technical Report WS-15-24, Crowdsourcing Breakthroughs for Language Technology Applications, pp. 1–8. The AAAI Press, Palo Alto, California (2016)
6. Dumitrache, A.: Crowdsourcing Disagreement for Collecting Semantic Annotation. In: European Semantic Web Conference, pp. 701–710. Springer, Cham, Switzerland. (2015)
7. Font, F., Roma, G., Serra, X.: Freesound Technical Demo. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 411–412. ACM, New York, NY, USA (2013)
8. Kairam, S., Heer, J.: Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing, pp. 1637–1648. ACM, New York, NY, USA (2016)
9. Kapelner, A., Kaliannan, K., Schwartz, H. A., Ungar, L., Foster, D.: New Insights from Coarse Word Sense Disambiguation in the Crowd. In: Proceedings of COLING 2012: Posters, pp. 539–548. The COLING 2012 Organizing Committee, Mumbai, India (2012)
10. Kittur, A., Chi, E. H., Suh, B.: Crowdsourcing User Studies with Mechanical Turk. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 453–456. ACM, New York, NY, USA (2008)

11. Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E. M., Shaw, A., Zimmerman, J., Lease, M., Horton, J. J.: The Future of Crowd Work. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 1301–1318. ACM, New York, NY, USA (2013)
12. van Miltenburg, E., Timmermans, B., Aroyo, L.: The VU Sound Corpus: Adding More Fine-grained Annotations to the Freesound Database. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 2124–2130. European Language Resources Association (ELRA), Paris, France (2016)

CaptureBias: Supporting Media Scholars with Ambiguity-Aware Bias Representation for News Videos

Markus de Jong[2], Panagiotis Mavridis[1], Lora Aroyo[3], Alessandro Bozzon[1], Jesse de Vos[5], Johan Oomen[5], Antoaneta Dimitrova[3], and Alec Badenoch[4]

[1] Vrije Universiteit Amsterdam, User-Centric Data Science Group
{lora.aroyo,m.a.dejong}@vu.nl
[2] TU Delft, Web Information Systems
{p.mavridis,a.bozzon}@tudelft.nl
[3] Leiden University
a.l.dimitrova@fgga.leidenuniv.nl
[4] Utrecht University
A.W.Badenoch@uu.nl
[5] Beel en Geluid
{joomen,jdvos}@beeldengeluid.nl

Abstract. In this project we explore the presence of ambiguity in textual and visual media and its influence on accurately understanding and capturing bias in news. We study this topic in the context of supporting media scholars and social scientists in their media analysis. Our focus lies on racial and gender bias as well as framing and the comparison of their manifestation across modalities, cultures and languages. In this paper we lay out a human in the loop approach to investigate the role of ambiguity in detection and interpretation of bias.

Keywords: Bias detection · bias in news video files · ambiguity-aware bias representation · disagreement · machine learning · crowdsourcing · human in the loop

1 Introduction

The interpretation of textual and visual media is typically a subjective process where personal views and biases are becoming interlaced with and indistinguishable from the actual media content. For example, ethnic groups can be misrepresented by numbers in crime reports [10] and international news agencies can adjust the contents of their reports to tap into certain biases that they believe are present in the intended public [9]. So, the different points of view typically get expressed as a disagreement among different authors and consumers of the media content. The disagreement can be seen as a signal to identify the presence of ambiguity and has an effect on the detection of bias in visual and textual media, as well as on the understanding the meaning of the media message.

Studies of visual and textual media bias can be quite labor-intensive when performed manually [21], *e.g.* through labeling manually hundreds of hours of video [9]. With the exponential growth of visual (news) content, many machine learning and human computation approaches are emerging for the automation of the labeling, analysis and processing of video and textual material. In this work, we aim at further extending the state of the art for large-scale processing of textual and visual media to support media professionals, humanities and social science scholars in their process of analyzing news media (with respect to studying framing, gender and racial bias in news). The central point here is the study of content and semantic ambiguity when it comes to determining the topic, the events and the sentiment of the media material. Further, we aim to understand what causes this ambiguity, what are different types of ambiguity and how they influence the understanding and the capturing of bias in visual and textual media across different languages.

The concrete objectives of this research are to support typical *digital humanities* analysis tasks, *e.g.*

- *distant reading* of large collections of visual and textual news for understanding patterns and contexts framing, racial and gender bias in news over time and across different cultures and languages
- *close reading* of specific instances of visual media for understanding aspects, properties and causes of framing, racial and gender bias in news over time and across different cultures and languages.

Therefore, we investigate the role of ambiguity of the media content, as well as the ambiguity of the topic(s), context(s) and specific event(s) and entities depicted in the news media for the detection of framing, racial and gender bias. Our research is guided by the following hypotheses:

- There are different causes for disagreement in interpretation of visual media that will lead to different types of ambiguity;
- Ambiguity found in visual media can be related to subjectivity;
- Different types of ambiguity and subjectivity can be used to detect different types of biases, such as framing, racial bias and gender bias.

2 Related work

Here we present the related work on disagreement and ambiguity that occurs after annotation tasks. As mentioned, disagreement is a signal for ambiguity or subjectivity. Then ambiguity itself can also be a sign of subjectivity. Then these signals appear in the different manifestations of bias through misrepresentation of *entities* with the method of *framing* [9] or with different *sentiments* attached to these entities. Some of the entities that contain gender and race can also often be misrepresented [18, 10]. In the following we present the work that is related to the the detection of the above signals and bias manifestations.

Methods that study or leverage the disagreement in order to identify the quality of annotations done by a crowd exist. For instance, in computational linguistics [4] use Generalizability theory as a means to capture the reliability of an annotation and identify the reasons behind the level of confidence and reliability we can have over an annotation. In [17] they also use crowdsourcing for annotations and identify different subgroups of disagreement between crowdworkers for annotations and compare them with expert annotations. Also [8] propose a different measure for agreement that solves a number of problems that arise when other agreement measures are used for interval values. Instead they propose to reason about the type of agreement or disagreement by looking into the distribution of answers within an interval of values when suitable for the problem. On the other hand, [25] identify also disagreement and divergence into groups of coders and evaluate two tree based ranking metrics to compare disagreements.

Crowdtruth is a platform [16] that applies disagreement analytics to generate ground truth data with the use of crowdsourcing. It has been used to identify and name entities as well as determine annotation ambiguity [15], to detect language ambiguity in medical relations in texts [11] and to determine intrinsic ambiguity of events in video event detection [14]. Another automated method that uses the crowd predicts the ambiguity of images to assist in an crowdbased foreground object segmentation task [13].

Now, we take a look at the types of bias we are interested in: framing, racial bias and gender bias. We give a short definition of these, followed by related research methods for those biases.

A *frame* of a message can be described as 'highlighting some bits of information about an item that is the subject of communication, thereby elevating them in salience' [12], and the act of *framing* can be described as 'selecting and highlighting some features of reality while omitting others' [12]. For research purposes, it is therefore important to find the amount of attention that is given to a certain element (*e.g.* highlighting or downplaying) and what is omitted.

Gender and racial bias in media is most often investigated via certain *misrepresentations* and *presentations* of groups. An example of misrepresentation is when the number of group X shown on screen is not representative of the number of group X that are part of that society. An example of difference in presentation is when group X is presented or described in an different manner, *e.g.* shown in different sentiment than group Y or described with different adjectives, or when the focus lies on different properties of the groups. Therefore, the goals for investigating gender and racial bias here are (1) quantitative comparison with population statistics for misrepresentation, and (2) the rather more complex qualitative comparison between groups of the representation.

Framing can be investigated through manual thematic analysis [21]. However, automated methods also exist such as using keyword clustering to identify stakeholders standing on different sides [19]. Word-based quantitative text analysis and computer assisted methods have also been used, *e.g.* to identify interest group frames in the framing of environmental policy in the EU [5]. In the case

of framing in video, we mentioned the investigation into framing in TV-news in countries that lie in overlapping spheres of influence of Russia and the EU [9], namely Belarus, Moldavia and Ukraine. In that study, 607 video news emissions were manually labeled on subject (EU, Russia), tone (positive, negative, neutral, none), theme (*e.g.* culture, history, security, values) and topic (*e.g.* external events or developments, human interest stories, visit from a state official). The relative number of reports on either EU or Russia was also compared. The results included statistics that showed different news channels aimed at particular local preferences (*e.g.* a shared religion, a shared history), but that (apart from the Russian channels) the news was in general most often balanced and neutral in tone and did not differ in tone towards either the EU or Russia.

As mentioned, research can discover racial bias expressed by discrepancies between actual on-screen role representation of ethnic groups and data from official statistics [10]. Example results from this 2017 investigation performed in Los Angeles showed that blacks were correctly reported as perpetrators, victims and police officers, and, while Latinos were accurately reported as perpetrators, they were underreported as victims and police officers. Whites were significantly overrepresented in all three categories. A similar quantitative comparison can be carried out to investigate gender bias, *e.g.* to investigate balanced reporting in sports [18]. This research also included qualitative research in which raters were asked to label announcer's language usage in relation to the athlete's gender (*e.g.* appearance, marital status) and imagery (*e.g.* active vs non-active pose, sports vs non-sports context). The researchers reported no significant quantitative gender bias, although there were still some differences found on other criteria. In other work, gender bias in Dutch newspapers expressed by stereotypical representation of male vs. female leadership in politicians was investigated with a dictionary approach [1].

To investigate framing and other biases, it is important to determine differences in message sentiment. Some automated text sentiment tools have been developed [20, 7] which are based on natural language processing (NLP). Voice tone is another possible source of sentiment analysis [26]. A relatively new modality in sentiment analysis is video, in which facial recognition techniques used to analyses actor's facial expression ('facial affect') [24]. Some work has also been done on creating an ensemble of all these sentiment analysis methods [22].

The methods put forward to analyze framing, gender and racial bias, however, do not make use of ambiguity in the crowd, even when such subjectivity may give us valuable information that could lead us to better detect bias and create better labels on subjective aspects as sentiment. Therefore, we propose an ambiguity-aware method that builds on CrowdTruth methodology [16] that will make use of ambiguity in the crowd to better detect bias.

3 The Approach: Disagreement-based Ambiguity for Bias Detection

We perform a number of knowledge acquisition experiments with media scholars and social scientists to determine aspects of bias in different modalities, cultures and languages. Next to this we also study ambiguity expressions, causes and types through crowdsourcing experiments for annotation of sentiment, topics, and opinions in news videos and articles. Main focus here is to understand (1) how disagreement is manifested as a signal for ambiguity, and (2) how ambiguity is related to subjectivity, and ultimately how these two lead to more accurate representation of bias in video and textual news. For this we apply, adapt and extend the CrowdTruth approach [3, 2, 16], which has been used to study disagreement-based ambiguity in various domains. We employ a hybrid human-machine system, where basic processing of both video and text material is performed to be used as a seed for the human computation tasks. Considering the large amount of video and text articles involved we envision an active learning cycle, where machine learning components continuously learn from humans-in-the-loop.

3.1 Dataset

Next, we describe the two types of data that we use and compare in our datasets: (1) textual and (2) video data.

Textual dataset Our textual dataset consists of news articles written in English from online sources such as: *e.g. BBC, The Guardian, CNN, Fox News, The New York Times, The Moscow Times, Sputnik, Breitbart News.* To identify target news events to study in videos, we use Wikipedia pages focusing on historical and political events [6]. Wikipedia provides crowd-sourced and editor-vetted articles from different contributors. We aim to extract event names and related event entities, e.g. people, organizations, locations and times and compare their representation in terms of opinions, perspectives and sentiment ground truth to compare the entities and facts presented within between different news sources.

Video dataset We perform experiments with a video dataset of short English language newsreels (*i.e.* a few minutes long with a spoken dialogue), accompanied by their metadata, *e.g.* short video description, title, tags, (auto-generated) subtitles and user comments. The videos in this dataset are collected from the following online news channels: *e.g. CNN, BBC, Al Jazeera, Sputnik, RT (formerly Russia Today), France24.* We also take advantage of the keyword annotated datasets on videos provided by YouTube in the YouTube8m dataset[7].

[6] Wikipedia: www.wikipedia.com
[7] YouTube-8M Dataset: https://research.google.com/youtube8m/

3.2 Data Preprocessing

We enrich the *subtitles*, *transcripts*, *in-video text* and *video metadata* with the set of events and related entities extracted from relevant Wikipedia pages and news articles.

Ambiguity signals in dataset We want to capture the different ambiguities from the dataset itself. For instance, using ControCurator[8] we process the comments from Wikipedia pages and YouTube videos from users in order to capture possible controversies. Also, for Wikipedia we can use a method similar to [23] in order to find controversial news articles from Wikipedia or Contropedia[9].

News event detection and data gathering After finding possible bias candidates with the use of the above tools from Wikipedia pages, we extract events using NLP processing. When Wikipedia articles are not present (for instance in the case of very recent news) we use different news article sources for the event and also make use of an initial video input from one source directly. We also use controversial video comments from these events, and, supported by Wordnet [10], we create seed words to assist a crowd to annotate an event. When the events are identified, we can collect video data from the different video channels of our initial dataset.

3.3 Disagreement for Bias Cues Extraction

In order to identify the framing, gender and racial bias introduced in news videos, we compare the information gathered from the video with Wikipedia and newspaper texts, as well as other videos (*e.g.* from other channels). When we are able to determine which main entities are related to an event, we can detect misrepresentions (of *e.g.* facts, actors) that might indicate framing. If a particular gender or race is misrepresented or represented in a certain way, we can infer gender and racial bias. As said, we base our bias cues on disagreement in both automatically extracted information and the crowd.

To be specific, order to be able to annotate videos for their events, we want to extract particular cues with both machine learning and human computation. Ideally, we want to identify with machine learning what needs to be annotated in the videos and transcripts by humans in order to find out *e.g.* what is being said, who is reporting, who is talking, how long are they talking, are they present at the scene of the news event?

To make use of all data modalities in our news videos, we investigate combining existing API's for textual, voice- and face-based sentiment analysis [22] in relation to the entities. Also, to be able to attach the entities to particular sentiments [6], we can compare different API's and state of the art methods and use their "disagreement" as a way to give a confidence to the combined output

[8] ControCurator: Crowds and Machines for Modeling and Discovering Controversy-http://controcurator.org/

[9] Contropedia: Analysis and visualization of controversies within Wikipedia articleshttp://contropedia.net/

[10] Wordnet: wordnet.princeton.edu/

and apply human computation to validate the sentiment analysis output from the machine learning methods. CrowdTruth[11] can be used to reason about the disagreement of the various subjects. Given that the crowd can also disagree for a particular subject, we investigate the reasons why the crowd could interpret a given message differently with regards to, for instance, their *demographics*.

4 Discussion

One of the limitations of our proposal is the lack of reliable data to capture 'opinion' neutral definition of recent events. As we use Wikipedia pages to extract both ground truth events to seed the search of these in media, as well as the intensity of edits and changes to these pages as an indication of possible controversy / bias or variety of opinions.

Acknowledgements

This research is supported by the Capture Bias project [12], part of the VWData Research Programme funded by the Startimpuls programme of the Dutch National Research Agenda, route "Value Creation through Responsible Access to and use of Big Data" (NWO 400.17.605/4174).

References

1. Aaldering, L., Van Der Pas, D.J.: Political leadership in the media: Gender bias in leader stereotypes during campaign and routine times. British Journal of Political Science p. 121 (2018). https://doi.org/10.1017/S0007123417000795
2. Aroyo, L., Welty, C.: The three sides of crowdtruth. Journal of Human Computation **1**, 31–34 (2014)
3. Aroyo, L., Welty, C.: Truth Is a Lie: CrowdTruth and the Seven Myths of Human Annotation. AI Magazine **36**(1), 15–24 (2015)
4. Bayerl, P.S., Paul, K.I.: Identifying sources of disagreement: Generalizability theory in manual annotation studies. Comput. Linguist. **33**(1), 3–8 (Mar 2007). https://doi.org/10.1162/coli.2007.33.1.3, http://dx.doi.org/10.1162/coli.2007.33.1.3
5. Boräng, F., Eising, R., Klüver, H., Mahoney, C., Naurin, D., Rasch, D., Rozbicka, P.: Identifying frames: A comparison of research methods. Interest Groups & Advocacy **3**(2), 188–201 (2014)
6. Calais Guerra, P.H., Veloso, A., Meira, Jr., W., Almeida, V.: From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 150–158. KDD '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2020408.2020438, http://doi.acm.org/10.1145/2020408.2020438

[11] CrowdTruth: The Framework for Crowdsourcing Ground Truth Data http://crowdtruth.org/

[12] https://capturebias.eu/

7. Chaumartin, F.R.: Upar7: A knowledge-based system for headline sentiment tagging. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 422–425. Association for Computational Linguistics (2007)
8. Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., Demartini, G.: Let's agree to disagree: Fixing agreement measures for crowdsourcing (October 2017), http://eprints.whiterose.ac.uk/122865/, © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org).
9. Dimitrova, A., Frear, M., Mazepus, H., Toshkov, D., Boroda, M., Chulitskaya, T., Grytsenko, O., Munteanu, I., Parvan, T., Ramasheuskaya, I.: The elements of russias soft power: Channels, tools, and actors promoting russian influence in the eastern partnership countries (2017)
10. Dixon, T.L.: Good guys are still always in white? positive change and continued misrepresentation of race and crime on local television news. Communication Research **44**(6), 775–792 (2017)
11. Dumitrache, A., Aroyo, L., Welty, C.: Crowdsourcing ground truth for medical relation extraction. arXiv preprint arXiv:1701.02185 (2017)
12. Entman, R.M.: Framing: Toward clarification of a fractured paradigm. Journal of communication **43**(4), 51–58 (1993)
13. Gurari, D., He, K., Xiong, B., Zhang, J., Sameki, M., Jain, S.D., Sclaroff, S., Betke, M., Grauman, K.: Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). International Journal of Computer Vision **126**(7), 714–730 (2018)
14. IEPSMA, R., GEVERS, T., INEL, O., AROYO, L.: Crowdsourcing for video event detection. In: Collective Intelligence (2017)
15. Inel, O., Aroyo, L.: Harnessing diversity in crowds and machines for better ner performance. In: European Semantic Web Conference. pp. 289–304. Springer (2017)
16. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., Sips, R.J.: Crowdtruth: Machine-human computation framework for sing disagreement in gathering annotated data. In: The Semantic Web–ISWC 2014, pp. 486–504. Springer (2014)
17. Kairam, S., Heer, J.: Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 1637–1648. CSCW '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2818048.2820016, http://doi.acm.org/10.1145/2818048.2820016
18. Kinnick, K.N.: Gender bias in newspaper profiles of 1996 olympic athletes: A content analysis of five major dailies. Women's Studies in Communication **21**(2), 212–237 (1998)
19. Miller, M.M.: Frame mapping and analysis of news coverage of contentious issues. Social science computer review **15**(4), 367–378 (1997)
20. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics. p. 271. Association for Computational Linguistics (2004)
21. Philo, G., Briant, E., Donald, P.: Bad news for refugees. Pluto Press (2018)
22. Poria, S., Peng, H., Hussain, A., Howard, N., Cambria, E.: Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing **261**, 217–230 (2017)

23. Rad, H.S., Barbosa, D.: Identifying controversial articles in wikipedia: A comparative study. In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. pp. 7:1–7:10. WikiSym '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2462932.2462942, http://doi.acm.org/10.1145/2462932.2462942
24. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. vol. 37, pp. 1113–1133. IEEE (2015)
25. Zade, H., Drouhard, M., Chinh, B., Gan, L., Aragon, C.: Conceptualizing disagreement in qualitative coding. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 159:1–159:11. CHI '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3173733, http://doi.acm.org/10.1145/3173574.3173733
26. Zhou, S., Jia, J., Wang, Q., Dong, Y., Yin, Y., Lei, K.: Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach (2018)

Bounding Ambiguity: Experiences with an Image Annotation System

Margaret Warren[1], Patrick Hayes[2]
[1] Metadata Authoring Systems
[2] Florida IHMC,

Abstract. This paper reports on a web interface for creating and editing rich metadata descriptions for images using RDF triples. We discuss the roles of ambiguity, disagreement and subjectivity in knowledge formation arising from an ongoing experiment in which users create semantic annotation of images, and we discuss the ways these elements have influenced the design of the system.

1 Background

ImageSnippets (http://www.imagesnippets.com) is a web interface designed for the construction of machine-computable image descriptions by lay users. Prompted by image cues, image annotators and/or subject matter experts can disambiguate entities from public corpora or define new terms through the interface. The subsequent captured knowledge is then stored in triple based graphs for inference, findability and reuse by semantically aware processes. The output formats conform to W3C linked data standards: (RDFa, JSON-LD) and utilize terms from the large and growing web-based concept datasets including DBpedia, YAGO and the Art & Architecture Thesaurus, without requiring users to be aware of this machinery.

The system has a number of features for capturing the highly subjective things people might say about images, but it also allows them to disambiguate their meanings in a suitably precise way, thereby providing a process for formalizing intuitive and expert knowledge. In our research with the system, we have found disagreements about meanings to be a centrally important methodological tool to create more useful and intuitive conceptual distinctions and that ambiguity is unavoidable and often useful, and should be considered something to be controlled and bounded appropriately, rather than eliminated; and that shared subjectivity is more important than objective truth or correctness when using semantic markup for inference and retrieval. Our work originally began as an exploration of how to bring together precise machine-readable descriptions in Semantic Web notations such as RDF and OWL with the informal language of how people actually describe images, especially (but not exclusively) the language of artists and deliberately considered the artists themselves as the subject matter experts on their own work (Eskridge et al, 2006, Warren & Hayes 2007, 2010). But this early knowledge representation work also had a secondary goal: that we could use our work to design a system with a (hopefully) intuitive interface that would capture this highly subjective, personalized

and informal knowledge in such a way that the resulting precise and disambiguated metadata could improve findability and then be distributed on the web with as much readability, interoperability and persistence as possible.

So the ImageSnippets system was designed first and foremost as a research tool to observe how users or image annotators would interact with the system, but we also considered that the metadata created through this research could also be accessed by any semantically aware processes and search engines and further, that the data could be properly attributed with provenance and publishing authority.

At the core of the ImageSnippets system is an intentionally small, lightweight ontology which provides a basic vocabulary of properties used to relate an image - the subject of the descriptions (or a region in the image) - to object values as entities which are then selected from public data sets and which, in other image annotation systems, would normally be referred to as keywords or tags. The annotators are given these properties as intuitive guidelines and encouraged to use terms from LIO first, but with some training, they can create new properties as well. The ability for users to create new properties is a highly useful way for the system to capture and analyze subtle distinctions and use these distinctions to build both new, domain specific ontologies and to extend LIO to allow annotators to talk, not just about images, but about the things that are seen in the images. LIO has evolved over many thousands of hours of use and testing, and the history of this evolution provides some suggestive lessons in the use of contradiction and ambiguity to discover intuitive concepts.

The design of the main ImageSnippets interface by necessity had to be concerned with capturing highly subjective and ambiguous sentiments from the naturally expressed ways people describe images, but incorporated in the design was also the creation of the core terms which make up LIO and this process was, in itself an iterative exercise in using something we call *bounded ambiguity*.

Outside of mathematics and some related pursuits, the meanings of words or symbols are always ultimately determined by the way that those words are actually used to pragmatically convey content, rather than by exact formalizable definitions. Even though Web protocols give IRIs an unprecedented degree of global exactness when used to retrieve content, this does not apply to their use as names in the Semantic Web, where their meanings are just as socially defined as words in natural languages. Widely cited calls to remove all ambiguity from IRIs (such as https://www.w3.org/wiki/GoodURIs) are doomed to failure.

Nevertheless, while ambiguity cannot be eliminated, it can be bounded. Published OWL ontologies and thoroughly documented public catalogs of IRI meanings such as DBPedia can both give IRIs a tighter. more exactly restricted meaning than is commonly found in natural-language words, and sometimes can adequately capture useful meaning distinctions which have not (yet) been adopted in normal English. Much of the art in designing useful Web data seems to involve locating the most useful bounds on ambiguity and we believe that bounded ambiguity has yielded some highly useful and encouraging search results.

Naturalness of use, and plausibility of the resulting descriptions, have been central to the goal of our project, which aimed to take image description to a level impossible to reach by the use of simple ‘keywords’ or ‘tags’. Image tags or keywords cannot overcome the sometimes extreme lexical ambiguity of English words in isolation, and cannot record how the indicated idea is related to the image. It should be noted here

that there are many existing annotation systems and crowdsourcing efforts that employ both complex schemas and keyword disambiguation as well as the ability to group keywords into more useful machine processable conventions; however the subject of our research has always been to push the edges of keyword meaning in relationship to the image.

Therefore, ours is a very different kind of activity from conventional ontology engineering. It is mostly performed by people experienced in the use of the software but with no philosophical or technical background, and in some cases with a high-school level of education. It rarely strays into 'upper-level' decisions about how things are to be classified, beyond a very basic and shallow taxonomy of classes of things. It is driven by the immediate needs of describing things seen in images, and failure to say something is not an acceptable option, so the subject matter is open-ended. Naturalness is essential, but descriptions cannot go beyond the restricted grammatical patterns of RDF linked triples, so concepts must sometimes be invented to express complex ideas. When this is done, the results are subjected to a Darwinian process of selection: if a new idea – usually a new relation name – is found to be useful in other, subsequent, annotations, it is retained, and this re-use in itself provides a growing body of evidence illustrating its actual meaning. We believe that, in this way, a slowly growing corpus of useful concepts - mostly RDF properties - are being accumulated which seem to allow quite rich expressiveness within the confines of the simple RDF triple graph syntax.

One important idea in our methodology has been that of the information recording point, where a user has a clear thought about an image and should be able to express this naturally, in an annotation which is detailed and structured enough to transmit that intended meaning to other users downstream. The information recording point is the moment that the annotator has something to say and is ready to say it. Our aim is to provide a tool that can be used at that point, without the annotator being required to see or think about anything beyond ideas expressed in English words and a readable rendering of their content. (The interface provides this as a grammatically primitive but readable 'pidgin' pseudo-English.) Coupled with this is an idea borrowed from software training usually in corporate cultures, that of 'just-in-time learning'. In a way, it could be said that we are giving annotators 'just enough' structure and just enough guidance in the interface to choose from one of the core concepts at the very moment the annotator is deciding what they want to say. Most users of the system can be trained in the basic construction of triples using the core LIO concepts in just a few minutes and after they have gained sufficient experience in the basics, they can easily then be trained to extend the triples to allow even greater expressivity.

The inherent ambiguity of the core LIO vocabulary seems to be part of the reason for its success. If we had imposed very 'strict' meanings on these relations, making finer-grained distinctions based on very exact meanings, it would of necessity have been several orders of magnitude larger, greatly increasing the cost of development but also the cost of use, since users would be required to make very precise distinctions at the point of use, with low tolerance for error.

It is difficult to talk about the design of LIO without also discussing some of the interface design decisions, because the vocabulary was deliberately allowed to grow somewhat organically around a combination of user workflow and some key philosophical decisions as starting points. So we first created the editor to enable the

triples to be written and almost immediately designed a complementary search and sort function to test the effectiveness of the properties. Additionally, as part of building the triples, users were provided with a look-up for finding and disambiguating the object values of the triples from public datasets. Choosing entities from sources such DBpedia, Yago, The Art & Architecture Thesaurus is also a highly subjective process which itself influences the choices made for properties.

To date, the basic LIO ontology consists of eleven properties (including a small number used to classify images into categories, such as being a photograph or a pencil drawing, and placing them in collections) have been found to be sufficient, together with the concepts from DBPedia and other online corpora, to create rich and useful descriptions of a wide variety of images.

2 Building LIO

Almost every aspect of the LIO ontology has its origin in a *disagreement*, either between ourselves or between us and some of our users. Sometimes these were resolved by careful disambiguation; but more often, by realizing that we were all using words differently, leading to the creation of a new concept name, whose true meaning was often decided by the patterns of its actual use. Some examples follow.

2.1 Images vs Works

The root subject IRI of all our descriptions identifies the digital image shown to the user as a thumbnail, but exactly what this IRI is supposed to denote in our descriptions is not entirely obvious, and was heavily debated. For a digital photograph the root subject is simply the image itself, but for a photograph of a painting in a museum catalog it is natural to say that 'the image' refers to the painting, and in some cases –for example, a cropped image of a sculpture in a museum catalog – it may even be a physical object. In general, we use the terminology of a *work*, as in 'work of art' and the subject of our descriptions is always the work.

But this introduces an ambiguity: is our subject IRI referring to the image itself or to a work seen in the image? At an early stage of the design of our interface we resolved this ambiguity by providing a carefully described set of alternatives, but this proved to be unworkable, as it required users to carefully make artificial-seeming distinctions, and to create awkward descriptions. If one delves into these distinctions deeply, one quickly descends into an intellectual quicksand. Who exactly is the creator of what you see in a photo of graffiti? Most people would say the photographer, unless it is something recognizable by an artist like Bansky. In our experience, the way an image is cropped seems to have some bearing on who might be said to be the creator. In some cases, it is the Contact Volume Editor that checks all the pdfs. In such cases, the authors are not involved in the checking phase.

We simply assume that the subject of our descriptions is a work, without recording explicitly whether this means the image itself or something it illustrates. (Intuitive guidelines are that a cropped, sharp image of a single object shown in an idealized

studio-lighting format is a picture of the work, while a simple photographic image of a landscape is the work These distinctions are not recorded explicitly, so the image IRI might refer to itself or to a work it illustrates. This is of course the classical philosophical confusion of use versus mention. This is an ambiguity that we have decided to embrace rather than disambiguate, which has has not turned out to be a problem in practice. Users seem to follow the required discipline intuitively without needing to be trained or corrected. We think that there are probably good reasons why some famous philosophical confusions are widespread in human affairs, even though they have been noted by scholars for thousands of years. The time or effort they save, compared to more accurate but more pedantic descriptions, is enormous, while the ambiguity they embody is easily resolved in practice. In our case, the use of multiple sources of information in descriptions, including the use of a 'creator' field, or the Dublin Core *dc:creator* property, or both, seems adequate.

2.2 Depiction vs Showing

The most basic LIO property is *depicts.* We debated reusing the depiction properties used in FOAF (Brickley & Miller 2010). But in spite of the widely recognized best-practice utility of re-use, we decided to introduce and use *lio:depicts* as a new property on the grounds that almost all FOAF uses will be images of people, whereas LIO is intended to apply to a much wider range of images. While nothing in the published FOAF vocabulary description requires this restriction to people, it seems clear that in actual use, the domain of foaf:depiction is people rather than owl:Thing; and we believe that use determines actual meaning.

Real images often show many entities that are incidental to the main subject (if there is one). Are they all depicted? A disagreement about pictures which 'accidentally' reveal a celebrity, while being aimed at other subjects, led us to introduce a distinction between depicting and merely showing. We noted that other systems exist which introduce a concept of weighting how central each thing depicted is in a scene, so we introduced a property *lio:shows* to be used for noting things in images which are in a sense incidental or not the 'main subject'. The distinction between *lio:depicts* and *lio:shows* is vague and underdetermined, but it seems natural, and there is a high degree of agreement in the use of these relations by different users in all the cases we have tested. Also, people use these terms with very little training and they use them rapidly, in much the same amount of time as it would take to create bare keyword tags. Figure 1 illustrates some examples of how the distinction works. Subsequently, all of the other LIO properties seem to be split away from 'depiction'

where a visual ambiguity needs to be resolved.

Figure 1: Depicts, Shows, UsesPictorially

2.3 Location vs. Setting

Early uses of a draft of the ontology revealed a form of use that we had not anticipated. It can be illustrated by the example of a photograph of some newly prepared food on a table. The photographer had added the description

lio:shows dbpedia:Pensacola,_Florida

on the grounds, when asked, that the photo had been taken in Pensacola. Clearly he was not using our property in the way we had intended. When we explained his error, his response was to say that he felt the 'setting' of the photograph was important.

It is worth noting that the issue of location in image annotation is in and of itself highly ambiguous and probably hotly contested in many circles. Early on, we had added a static location field (re-using an Adobe XMP location that can often be found embedded in image data) to attempt to capture at least one version of location initially; but there are at least 4 notions of location: the location of the camera or input device that captured the digital image, the location where a work might have been created, the location of a scene in the work and the location of where the work may currently be located. In all of these cases, attempts have been made to classify the distinction between scene location versus image location — and even the names of the location fields themselves become ambiguous and tend to change over time.

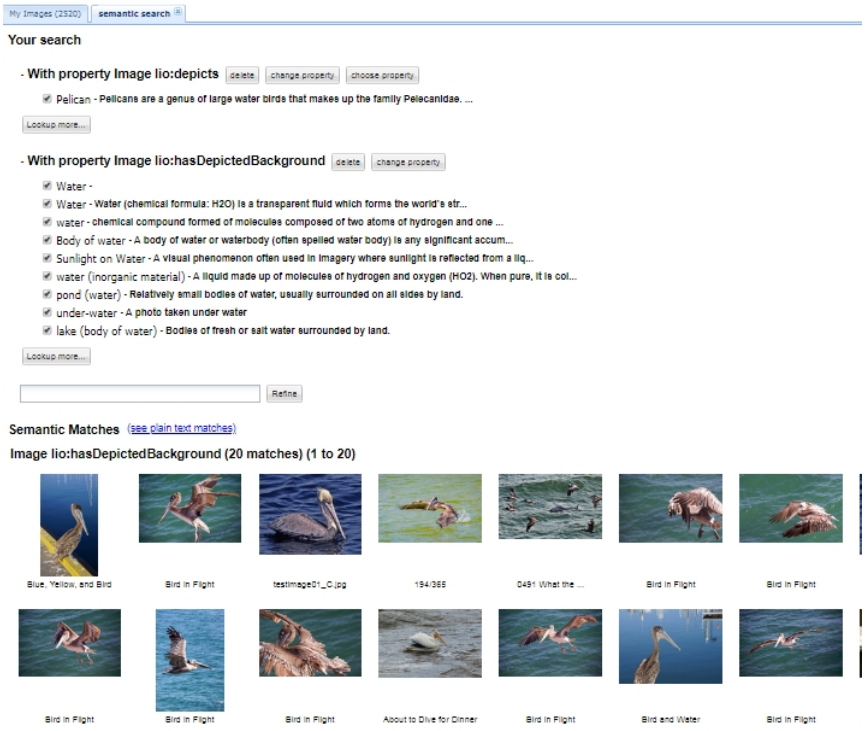
So we created a new property, *lio:hasSetting*, to relate an image to the broader place where it was created while still providing in other places in the interface, a way for annotators to deal with camera or scene locations. This judicious use of bounded ambiguity was rapidly overtaken by our users, who quickly extended its use further than we had imagined. They included events (the 1939 NY World's Fair), periods of time (the 17th century, when a portrait was painted) and even such things as 'kitchen', 'wedding', and even a person – for the setting of a tattoo. While these uses were not what we had in mind, they seem to work, and indeed we now find them natural. We therefore decided to simply allow all such uses, and rather than regard this as a situation where the range of the property is ambiguous, think rather that the users of LIO are, by their use of this property on thousands of images, creating a category, which might be called the class of Settings.

We have no way to exactly characterize this class in a formal ontology, and it does not occur in any published ePedia, but we can say that it has a large overlap with what one might call spatiotemporal envelopes, and includes a wide range of things, such as *twilight* and *dusk*, which are hard to classify. It need not be shown or depicted in the image itself (although it some cases it can be, as in a room or a park or someplace one might be 'in' when they make the image) And this is enough to make it a useful and intuitive element of the LIO vocabulary. The key point is that users find it natural to use, and they use it consistently; which we take as evidence that it corresponds to a widely shared concept.

2.4 Foreground and Two Backgrounds

The notions of foreground and background are of course familiar, and we provided relations *hasInForeground* and *hasInBackground* to allow users to talk of objects in the foreground or background. But we quickly discovered that there are two additional notions of 'background' that could be utilized with a fair amount of ease once an annotator had learned they could use the terms. First, there is the flat 2-D background field of a rendered image versus the 3-D background of a depicted scene, the stuff 'behind' the main object depicted in an image. We needed the former in order to state precise search criteria for works of art (a painting by Keith Haring with a white background) when most 'backgrounds' were distant areas in photographic images. In this case, a real (but unforeseen) ambiguity had to be bounded by providing distinct IRIs for two subtly different but visually important senses of an English word; even though the meanings of our current relations *lio:hasPictoriaBackground*, *lio:hasDepictedBackground* are (of course) themselves still somewhat ambiguous, we find that when used, they are used consistently, perhaps because the very presence of the distinction in the small vocabulary draws user's attention to fact that they can make this distinction explicit. *HasPictorialBackground* can refer to any kind of pattern or texture (such as 'checkerboard' or 'swirls') while *hasDepictedBackground* might be something like 'sand' behind a crab. These distinctions are valuable if perhaps one is searching for a pelican completely surrounded by water as opposed to sky (See Figure 2.)

Figure 2. Pelicans with Water as Depicted Background



2.5 Looking Like

Early in the LIO project we had a photograph of clouds and sunlight that some people said, depicted an 'angel'. This gave rise to a debate. Did it depict an angel? Some said, obviously yes, since this was the whole point of the photograph. Some said, obviously no, since there are is no actual angel to be depicted, only clouds, sky and sunlight. Some said angels did not exist and others emphatically said that medieval religious art was full of depictions of angels. We resolved this contradiction by introducing the notion of one thing looking like another. The image *depicts* clouds, but *lio:looksLike* an angel. Fortunately, the argument about the *existence* of angels was resolved by *http://dbpedia.org/resource/Angel*.

What seemed at first to be a simple conceptual hack for a limited class of 'illusions' turned out to be immediately useful in a large variety of descriptive situations, and *lio:looksLike* is now easily used by annotators. When reading plain-text descriptions of images, we often see phrases "looks like a bird" or "looks like a face". One might call this the visual equivalent of metaphoric language.

2.6 Conveys

Early in the project, marking up a blue period Picasso, someone wanted to say that it *shows* sadness. This produced immediate objections, such as where exactly was this shown in the image? Again, the resulting disagreement/discussion led to a new conceptual distinction, and the addition of *lio:conveys* to the vocabulary, intended to refer to emotions and moods that images do not in any sense show or depict. But just as with *lio:hasSetting*, *lio:conveys* gets used in ways we had not initially anticipated and yet make perfect sense. So, for example an image can *lio:conveys* emptiness, friendship or even a 'party' (as in a 'party atmosphere') and so we realized that annotators might also use it to refer to an idea. So far there seems to be no reason to further delineate this concept. By using the search interface in the system, we have

seen that the use of the ‘conveys’ property is basically never used to ‘convey’ something ‘depicted’.

2.7 Artistic considerations

The final two properties added to LIO were: hasArtisticElement and lio:usesPictorially. Throughout the design of the interface and the design of the ontology, we were using teams of annotators who would annotate small groups of images and then run multiple types of searches over the freshly created data so that we could examine the usage of both the vocabulary and the similarity in the object terms chosen from (mostly) DBpedia and Yago. During this testing phase, we began to observe patterns in the usage of the terms and realized that the teams were naturally aligning themselves to their own conventions on edge cases. Take, for example a photo of a guitar in which the guitar takes up a large area of the image, but is still clearly not the only thing in the image. Does the image ‘depict’ a guitar or ‘show’ a guitar? In the beginning we might have seen annotators choose ‘depicts’ until they ran a search and observed that there was clearly a visible difference in the search results and thus, they aligned themselves using the images themselves as guides. While these testing sessions were occurring, we additionally discovered two other properties that were eagerly embraced, one was for the notion of an Artistic Element. We began searching for terms like: spirals, movement, shadow and blur; words the annotators had been using, and realized that a category of general concepts related to classical art theory design elements: line, form, shape and color attributes could be useful. Once the property existed, annotators quickly began using it to describe things like ‘vanishing points’ and ‘symmetry’.

The last property added was one called: *lio:UsesPictorially*. In practical use, while annotators were “triple-tagging” images and then testing their use of the properties, another distinction between objects was observed in the search results that could not be handled by any of the other properties. The distinction was a subtle difference in the way one would say they ‘normally’ see something real in the world depicted and then those items that seem to be visually represented in an unusual way – perhaps an artistic sense, or maybe just photographed in an unconventional way. This distinction most often includes artistic renderings of objects in the form of paintings or illustrations, but it often goes beyond that into abstract photos of everyday objects or unusual angles. Almost every new annotator learns this property very quickly and though there can be ambiguous edge cases here as well, this property has seemed to bear a great deal of utility in search results. Figure 1 illustrates the distinctions between depicts, shows and usesPictorially in images of guitars

3. Describing what is depicted

While working on the design of LIO, we were also working on the design of the system itself. The core of the ImageSnippets interface is an editor window called the ‘triple editor’ in which most of the user activity occurs. Over the years we have experimented with several design options and consider this part of the system a work in progress.

Our goal has been to make the process of finding, and sometimes of coining, concept IRIs as quick and as painless as possible and at the moment, we have multiple ways of accomplishing the creation of the triples including ways to select multiple images at one time to annotate all of them with the same triple(s) simultaneously as well as entity extraction from free text input provided by subject-matter experts (SMEs) and direct image-to-word suggestions made by Clarifai and CloudSight, as shown in figure 3.

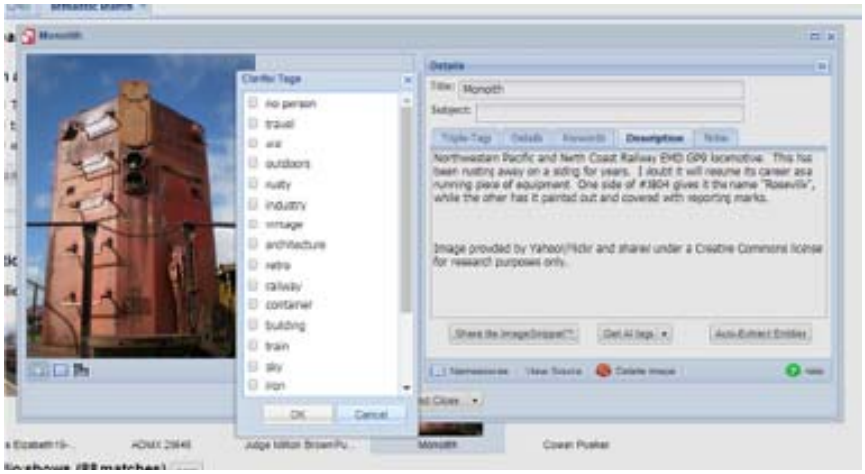


Figure 3: Clarifai (AI generated keyword suggestions) and the free text description by the SME

The basic creation of a triple involves the user typing in a word or phrase. At this point, we have a large number of concepts that annotators have previously found in all of the public corpora we have used to date. These concepts are auto-completed and presented to the user who can hover over the concepts to make sure they are disambiguating the term correctly. This auto-complete function in itself aids in the alignment of ambiguous results, since it is natural for a user to select an entity that has previously been chosen if it matches their intended sense of the concept they are describing. However, if it is a new concept – not previously used in the system, or if the user chooses – they can go to a full lookup and scan through a look-up window that shows all of the possible entities that might match the term the user has typed to find a correct definition that matches the sense of the word they had in mind. This can often be the most time consuming stage of creating the triples and one of the most challenging aspects of the design of the system.

It was our original intent to use just one public dataset to look up terms; however, we quickly found that not anyone dataset by itself contained enough range of concepts to allow for the expressiveness of real annotation. So, while the system allows both custom datasets to be used and the creation of new datasets; the default look-up

displays results from DBpedia, Yago, Art and Architecture Thesaurus and Wikidata. This, in and of itself can lead to duplications in triples across different datasets, but also occasionally forces annotators to make arbitrary choices among needlessly fine distinctions of meaning. The word 'curve' for example yields two senses in Yago: a curved section of a road or railway (the object itself) or the property of something (such as a roadway) being curved. This excess of conceptual distinctions based on fine philosophical or grammatical considerations or simple due to duplication is a source of difficulty for our project. Still, we cannot 'not' show all results, because in some cases – that fine grained distinction *is* necessary.

While the various corpora have many *owl:sameAs* links identifying similar concepts, these are not useful in practice. They are unreliable; but in any case, reasoning with equality is expensive and slow. What is needed is to have a single source of canonical names for concepts at just the right level of ambiguity and for now, we have made some headway in resolving this sometimes immense 'embarrassment of concepts' by hand over the course of hundreds of person-hours to build the cached concepts found by auto-completion. We also use some programmatic tricks that organize the results in the look-up such that the most likely matches will be at the top of each dataset. In any case, over time, power annotators quickly learn the fastest paths to disambiguation and this leads to increasingly more fluid look-up overall.

4. Visual Ontology Building

The development of the system has also given us multiple opportunities for experimentation in ontology development. First, we added the ability for the annotators to build more complex descriptions by chaining triples together by use of a 'which' and 'and' statements. In this way, more skilled annotators can not only describe the objects seen in the image, but they are also creating a slowly growing corpus of useful concepts - mostly RDF properties which seem to allow quite rich expressiveness within the confines of the simple RDF triple graph syntax. As images can depict just about anything, even mythological or illusory things, this process often pushes the edge of the expressive abilities of the published vocabularies, requiring power users to create new properties as well as new nominal concepts. Figure 4 illustrates an example of triples chained together with new 'properties' which are being accumulated in the system for further analysis. Often these are concepts arising in specialized domains, such as *cookingMethod* and *FoodStage* used to describe images illustrating recipes, *isWearing* used to talk about people in an image, and *hasVIN* to identify particular cars. But some are of more general utility, such as *isPartOf*, *isMadeOf,* and *hasColor*. Some of these have become extended in use, just as in the case of *lio:hasSetting*. One in particular, *hasCondition*, was introduced for describing stages in a process, such as a building *under construction* or a car engine *being assembled*, but its use has extended to things like *a curved road*, made up from the available concepts of *dbpedia:Road* and *yago:curve*. It can be used now to indicate a complex concept in which one descriptor modifies another. In many ways this seems to capture the adjectival construction in English, while staying within the simple

triple-graph syntax of RDF. This process of Darwinian selection (by frequency of re-use) and conceptual generalization is evolving a general-purpose conceptual language for describing things, which builds on and extends the basic semantic corpora; and is of course itself, uniquely, documented and illustrated by the images themselves.

Another style of usage for ontology development has also been found especially productive in which subject-matter experts (SMEs) provide image descriptions in free text, which is then processed by trained ImageSnippets power annotators who create the actual triple, using multiple techniques for more rapid triple building including the use of an auto-entity extractor from DBpedia over the free text which users can then automatically add to a list of pre-disambiguated concepts for attaching to LIO or their own user-created properties.

In cases like this, the concepts used by the SME are often more specialized, reflecting their expertise, than those found in the conceptual corpus. For example, a vehicle may be identified as a rare 1953 Fletcher prototype (one of only two still in existence) rather than simply a military vehicle, which is the best suggestion made by

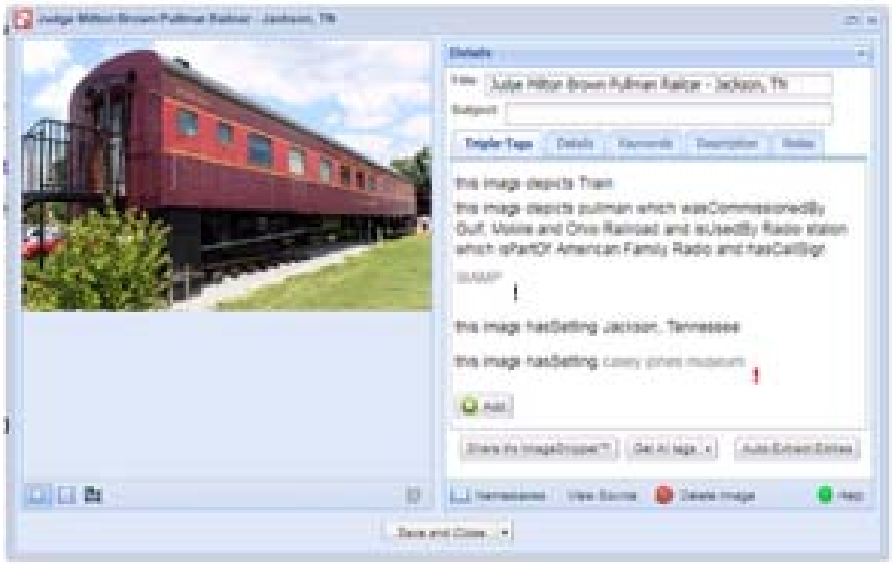


Figure 4. Use of chained triples in the Triple Editor

Clarifai. In these cases, the annotator can generate new IRI's for such terms, while noting and recording the subclass or instance relationships between these and the

existing terms in the public corpus. In this way, a specialized ontological vocabulary can be created as a byproduct of the image annotation process.

5. Conclusion

In conclusion, our long-term image annotation project, predominantly a labor of love by the first author, has had a number of goals, the first and foremost of which has been the construction of a linked-data, image annotation 'sandbox' for extended research and development in the areas of knowledge capture and representation using images as stimulus. Secondary goals involve the more practical use of sharing, managing and publishing the images and their metadata as well as an interest in using the system for the curation and preservation of historically significant digital image assets, particularly those from niche or at-risk domains with extremely deep and precise metadata. Much of our work has involved highly subjective and ambiguous distinctions, trial and error, testing and re-testing while simultaneously making interface changes to improve usability and the training of annotators to work in unfamiliar mental territory. Our methodology, in favor of intuitive usability has embraced a process of creating boundaries around ambiguity eschewing rigid definitions. Through a process of debating and resolving disagreements by splitting concepts, we attempt to identify useful conceptual generalizations as they emerge and adjust the boundaries of natural meanings. This is work in progress, but we are optimistic.

6. References

Eskridge, T., Hoffman,R., Hayes, P. and Warren, M. 2006 Formalizing the informal: a Confluence of Concept Mapping and the Semantic Web. *Proc. Second International Conference on Concept Mapping*, A J Cañas & J. Novak, eds. Costa Rica, 2006

Warren, M., and Hayes, P. 2007 Artspeak: The Contemporary Artist Meets the Semantic Web. Creating Formal Semantic Web Ontologies from the Language of Artists *Electronic Techtonics - Thinking at the Interface;* HASTAC (Humanities, Arts, Science, and Technology Advanced Collaboratory) - Duke University, 2007

Brickley, D. and Miller, L. 2010 FOAF Vocabulary Specification 0.97 http://xmlns.com/foaf/spec/20100101.html

Warren, M, 2016 A Visual Guide to the ImageSnippets Properties. http://www.imagesnippets.com/ArtSpeak/help/properties.html

Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis*

Mike Schaekermann[1], Edith Law[1], Kate Larson[2], and Andrew Lim[3]

[1] HCI Lab, School of Computer Science, University of Waterloo
[2] Artificial Intelligence Group, School of Computer Science, University of Waterloo
[3] Division of Neurology, Sunnybrook Health Sciences Centre, University of Toronto

Abstract. Low inter-rater agreement is typical in various expert domains that rely in part on subjective evaluation criteria. Prior work has predominantly focused on expert disagreement with respect to individual cases in isolation. In this work, we report results from a case study on expert disagreement in sequential labeling tasks where the interpretation of one case can affect the interpretation of subsequent or previous cases. Three board-certified sleep technologists participated in face-to-face adjudication sessions to resolve disagreement in the context of sleep stage classification. We collected 1,920 independent scoring decisions from each expert on the same dataset of eight 2-hour long multimodal medical time series recordings. From all disagreement cases (29% of the dataset), a representative subset of 30 cases was selected for adjudication and expert discussions were analyzed for sources of disagreement. We present our findings from this case study and discuss future application scenarios of expert discussions for the training of non-expert crowdworkers.

Keywords: Inter-rater disagreement · Adjudication · Sequence data.

1 Introduction

One of the most common use cases for crowdsourcing is the classification of objects into categories. While crowdsourced classification tasks traditionally focused on problems not requiring domain expertise, recent work suggests that crowdsourcing can also be effective for expert-level classification. Examples of such expert tasks from the medical domain include the identification of low-level patterns in sleep-related biosignals [22], the annotation of retinal images [12], and medical relation extraction [3].

In many mission-critical expert domains including the interpretation of medical data, low inter-rater agreement rates are the norm [5, 11, 15, 16]. Expert disagreement, however, poses fundamental challenges to quality control procedures in crowdsourcing, and to the use of data labels in supervised machine

* Supported by NSERC CHRP (CHRP 478468-15) and CIHR CHRP (CPG-140200).

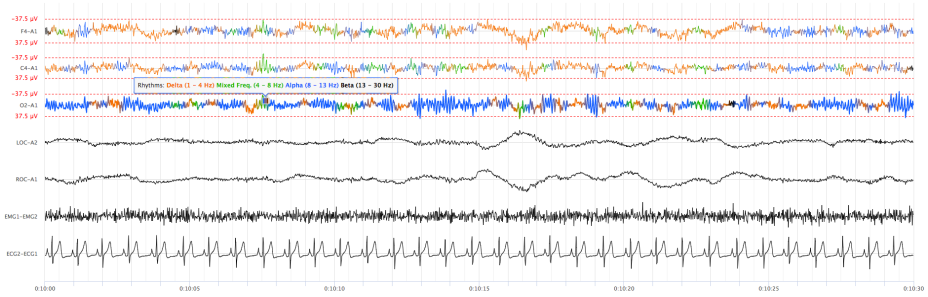


Fig. 1. Visualization of one 30-second epoch of biosignal data to be scored into one of five sleep stages

learning, as it is not immediately obvious how cases at the inter-subjective decision boundary should be disambiguated if multiple equally-qualified domain experts exhibit genuine disagreement.

Prior work has predominantly paid attention to the nature, sources and resolvability of expert disagreement on individual classification tasks in isolation [1, 4, 13, 21]. Many interpretation tasks, however, are sequential in nature, i.e., the interpretation of one case affects the interpretation of subsequent or previous cases. For example, in text translation, the semantic interpretation of one phrase or sentence can affect the translation of subsequent or previous phrases or sentences. Heidegger called this reciprocity of text and context the *hermeneutic circle*. Overall, sequential labeling makes up a large and diverse class of problems from numerous expert domains.

In this work, we present findings from a case study on expert disagreement in the context of sleep stage classification, the expert task of mapping a sequence of fixed-length pages of continuous multimodal medical time series (*polysomnogram*, see Figure 1) to a sequence of discrete sleep stages (*hypnogram*). Prior work has established that inter-rater agreement in sleep staging averages around 82.6% [17]. The objective of this case study is to identify various sources of expert disagreement in sleep stage classification and to investigate if and to what extent disagreement may be specific to the sequential nature of the labeling task and underlying data.

To answer these questions, we collected 1,920 independent sleep scoring decisions from a committee of three board-certified sleep technologists. We then selected a representative subset of the resulting disagreement cases which were resolved through in-person adjudication among the members of the expert committee. The rest of this paper describes the related work, then details our study for collecting and analyzing the expert deliberation data, and concludes with a discussion of application scenarios for the training of non-expert crowdworkers.

2 Related Work

2.1 Ambiguity and Sources of Inter-rater Disagreement

Ambiguity, the *quality of being open to more than one interpretation*, and the phenomenon of expert disagreement are central to the justification of knowledge, and have been extensively discussed in the epistemic literature [1, 4, 13, 21]. An early theoretical investigation named three types of expert disagreement [13]: *personality-based* disagreement arising from the incompetence, ideology, or venality of experts, *judgment-based disagreement* arising from information gaps, or *structural disagreement* that arises because experts adopt different organizing principles or problem definitions. Garbayo [4], on the other hand, distinguished a form of *legitimate* disagreement, that arises when experts can access the same evidence, but still diverge in interpretations, from *verbal* disagreement, i.e., misunderstanding among experts due to discrepancies in terminology.

Recent work in the field of human-computer interaction (HCI) has explored the issue of disagreement in the context of crowdsourcing tasks. Gurari and Graumen [6] analyzed visual question answering tasks and found that disagreement can be attributed to ambiguous and subjective questions, insufficient or ambiguous visual evidence, differing levels of annotator expertise, and vocabulary mismatch. Chang et al. [2] proposed to elicit help from the crowd for refinement of category definitions, based on the finding that workers may disagree because of incomplete or ambiguous classification guidelines. Kairam and Heer [8] introduced a technique to identify clusters of workers with diverging, but legitimate interpretations of the same task. Their work shows that disagreement can arise from differences in how liberally or conservatively workers interpret classificaion guidelines.

Our study revolves around the task of biomedical time series classification, a field with typically low inter-scorer reliability. For example, Rosenberg and van Hout [17] conducted a large-scale study on inter-scorer reliability in sleep stage classification and found that average expert agreement is as low as 82.6%. In a comment on this study, Penzel et al. [15] explained that systematic studies on the inter-rater reliability of sleep automatically bring up the question of truth, claiming that the "true" state (i.e., sleep stage) is unknown and can only be approximated through aggregation of expert opinions.

2.2 Group Deliberation as a Method for Disambiguation

Group deliberation is an interactive form of decision making among humans which typically involves group members with conflicting beliefs who try to reach consensus on a given question by presenting arguments, weighing evidence and reconsidering individual positions.

Several works explored factors that affect the process and outcomes of group deliberation. Solomon [20] appreciates conflict as an important phenomenon of any fruitful deliberation process. He argues that dissent is both required and useful—as *"dissenting positions are associated with particular data or insights*

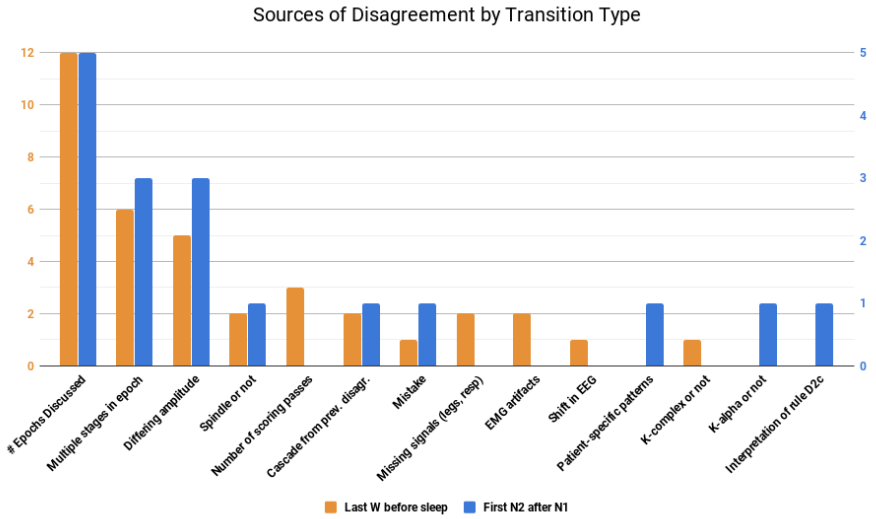


Fig. 2. Sources of disagreement by transition type. The vertical axis plots the number of times a particular source of disagreement was mentioned in an expert discussion about a case from one of two transition types: Last Wake before sleep onset and transitions from N1 sleep to N2 sleep. Note the two vertical axes, one for each transition type, are re-scaled to facilitate a visual comparison of both distributions relative to the number of epochs discussed (# Epochs Discussed) for each transition type. Expert discussions could mention more than one source of disagreement.

that would be otherwise lost in consensus formation"—and criticizes procedures endowed with the a priori aim of reaching consensus. Instead, he advocates for a structured deliberation procedure that avoids the undesired effects of *groupthink* [7] by actively encouraging dissent, organizing individual subgroups to deliberate on the same question, and ensuring diverse group compositions.

Kiesler and Sproull [9] found that time limits imposed on deliberation tend to polarize discussions and to decrease the number of arguments exchanged. The same work suggests the use of voting techniques or explicit decision protocols to structure the deliberation process.

Recent work by Schaekermann et al. [19] introduced a real-time deliberation framework to disambiguate edge cases in crowdsourced classification tasks drawing inspiration from some of these early design considerations. The same work also introduced a novel public deliberation dataset including all deliberation dialogues, original and revised classification decisions, and evidence regions from two different text classification tasks.

Navajas et al. [14] studied the effectiveness of in-person group deliberation for general-knowledge questions reporting that averaging consensus decisions yielded better results than averaging individual responses.

2.3 Consensus Scoring in Medical Data Analysis

Group deliberation has also been proposed as a technique for disambiguating edge cases in the interpretation of medical data. Rajpurkar et al. [16] employed group deliberation among cardiologists to generate a high-quality validation data set in the context of arrythmia detection from electrocardiograms (ECGs). Their work revealed that a convolutional neural network trained on independent labels (i.e., labels collected without deliberation) exceeded the classification performance of individual cardiologists when benchmarked against the consensus validation set.

Krause et al. [11] compared majority vote to in-person deliberation as techniques for aggregating expert opinions for diagnosing eye diseases from photos of the eyeground. Compared to majority vote, in-person deliberation yielded substantially higher recall, suggesting the potential of group deliberation for mitigating underdiagnosis of diabetic retinopathy and diabetic macular edema. Krause et al. also showed that performing group deliberation on a small portion of the entire data set can make tuning of hyperparameters for deep learning models more effective. The same consensus data set was later used by Guan et al. [5] to validate the classification performance of a novel machine learning approach involving the training of multiple grader-specific models. They demonstrated that training and aggregating separate grader-specific models can be more effective than training a single prediction model on majority labels.

In the context of sleep stage classification, Penzel et al. [15] refer to the concept of in-person group deliberation as *consensus scoring*, concluding that an *"optimal training for [...] sleep scorers is participation in consensus scoring rounds"*. In this work, we translate this idea to the non-expert domain suggesting a method to augment training procedures for crowdworkers through the use of edge-case examples and the associated expert discussion dialogues in the context of sleep stage classification.

3 Expert Deliberation Data Set

An in-person deliberation study was conducted with an expert committee of three board-certified sleep technologists at Sunnybrook Health Sciences Centre in Toronto to investigate the extent and potential sources of inter-rater disagreement in sleep stage classification, and the effectiveness of group deliberation as a method for consensus formation.

3.1 Data Set

We prepared a data set of eight 2-hour-long PSG recording fragments. Each 2-hour-long fragment contained a sequence of 240 30-second epochs of biosignal data, resulting in 1,920 (240 x 8) epochs for the entire data set. Half of the fragments were from healthy subjects, the other half from patients with Parkinson's disease. Both parts of the data set (Healthy and Parkinson) contained

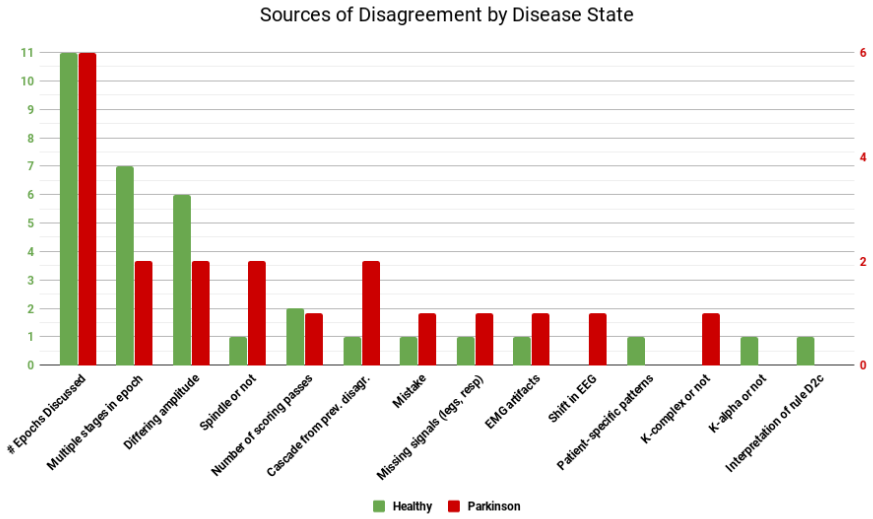


Fig. 3. Sources of disagreement by disease state. The vertical axis plots the number of times a particular source of disagreement was mentioned in an expert discussion about a case from one of two disease states: Healthy and Parkinson's Disease. Note the two vertical axes, one for each disease state, are re-scaled to facilitate a visual comparison of both distributions relative to the number of epochs discussed (# Epochs Discussed) for each disease state. Expert discussions could mention more than one source of disagreement.

examples of different transition types. We included examples from four different transition types identified by Rosenberg et al. [17] as regions with typically low inter-rater agreement: the last epoch of stage Wake before sleep onset, the first epoch of stage N2 after stage N1, the first epoch of stage REM after stage N2, and transitions between stages N2 and N3.

3.2 Procedure

The full data set was first scored independently by each sleep technologist, resulting in 5,760 individual scoring decisions, three for each of the 1,920 epochs. We then identified all epochs with disagreement among scorers and selected a subset of 30 epochs for in-person group deliberation. The selected disagreement epochs represented both disease states and all four transition types. All 30 epochs were discussed in person by the three scorers using a graphical scoring interface to facilitate detailed discussions about patterns present in the time series data. The experts participants were not explicitly required to reach unanimous consensus, and could instead choose to declare a case as *irresolvable*. We did not impose an explicit voting scheme or limit the amount of time available per discussion, but instead left the discussion dynamics open until all experts either agreed on one

	Tech A	Tech B	Tech C	Majority	# Obs.
Tech B	0.71	—	—	—	
Tech C	0.71	0.68	—	—	N=1920
Majority	0.87	0.83	0.84	—	
Deliberation	0.63	0.50	0.02	**0.54**	N=30

Table 1. Pairwise agreement between all sleep technologists (Tech A, Tech B, Tech C), as well as the group labels as determined by majority vote and the deliberation process. Agreement is measured by Cohen's kappa.

sleep stage or declared a case as irresolvable. Unanimous decisions were reached for all 30 epochs through a process of verbal argumentation and re-interpretation of the patterns shown in the biosignal data The *irresolvable* option was never used. Discussions were recorded (screen capture and audio), transcribed and qualitatively coded for the different sources of disagreement.

3.3 Inter-rater Disagreement

We measured pairwise agreement between all scorers (Tech A, Tech B, Tech C), as well as the group labels as determined by majority vote (Majority) and the deliberation process (Deliberation). Agreement was measured by Cohen's kappa. Table 1 summarizes all agreement results. Pairwise agreement among scorers was moderate, ranging between 0.68 and 0.71 (N=1920). Agreement between individual scorers and the majority vote was high, between 0.84 and 0.87 (N=1920). For the epochs discussed in person, we measured pairwise agreement between the deliberation decision and individual scorers' decisions. Two of the three scorers showed weak agreement with deliberation outcomes (Cohen's kappa of 0.63 and 0.50, N=30), while the third scorer showed no systematic agreement with the deliberation outcomes (Cohen's kappa of 0.02, N=30). Agreement between the majority vote and deliberation decisions was low (Cohen's kappa of 0.54, N=30).

3.4 Sources of Disagreement

Initial qualititative coding of the expert discussions for 17 cases from two major transition types revealed a broad range of reasons why sleep technologists may disagree on the correct sleep stage label. Figures 2 and 3 compare the relative frequency of different sources of disagreement across two transition types (*Last W before sleep* and *First N2 after N1*) and across two disease states (*Healthy* and *Parkinson*) respectively. Overall, we identified two sources of disagreement which occurred with the highest frequency in both transition types and disease states. These were (a) the presence of multiple stages in one epoch causing disagreement about which stage was the dominant one, and (b) different configurations of the graphical scoring interface in terms of amplitude scaling causing divergent interpretations of visual patterns in the signal.

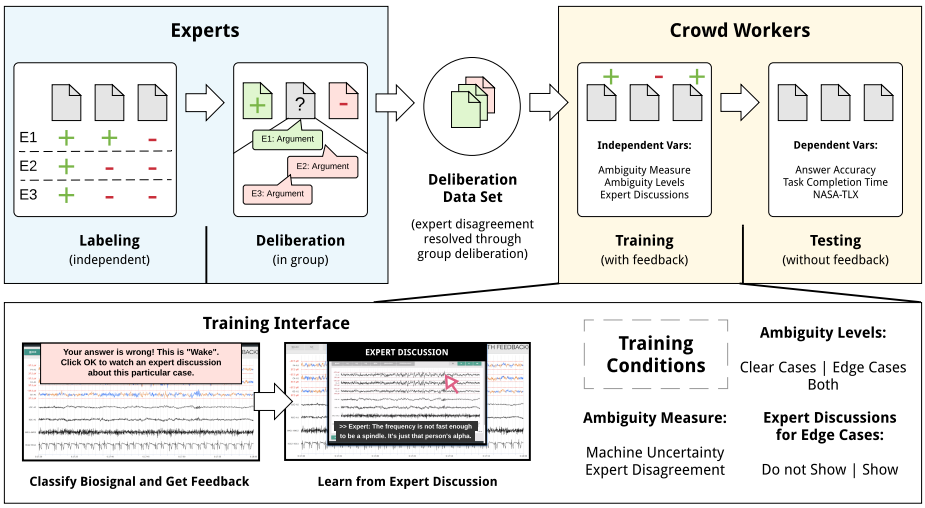


Fig. 4. Application scenario of using expert discussions for improving example-based training for non-expert crowdworkers.

While these two sources of disagreement could persist on individual cases without the sequential context, we identified two other sources of disagreement that explicitly depend on the sequential nature of the labeling task and underlying data:

- **Number of scoring passes**:
  for 3 out of 30 adjudicated cases, experts explicitly mentioned that their scoring decision depended on the number of passes they had taken on a particular recording. In other words, experts indicated that their interpretation of biosignals is often updated once certain patient-specific patterns are observed towards the end of the recording. A subsequent re-interpretation (i.e., second scoring pass) would then allow experts to take into account observations they have made in the other parts of the data sequence in one of the earlier scoring passes. Disagreement could therefore arise if one expert had only performed one initial pass whereas other experts may have performed two or more passes.
- **Cascade from previous disagreement**:
  3 out of 30 adjudicated cases could be resolved automatically once the disagreement on one of the close-by preceding cases had been resolved. This dynamic was observed since evidence for specific stages of sleep may sometimes be observed only at the transition point from one sleep stage to another. Consequently, disagreement may arise at a "critical" transition point and persist over multiple steps in the sequence. Once the disagreement at the transition point is resolved, the resolution can cascade to the subsequent steps until the next transition point.

These two sources of disagreement co-occurred once, meaning that 5 out of 30 adjudicated cases (17%) were associated with sources of disagreement that depend on the sequential nature of the labeling task and underlying data.

4 Discussion

In this work, we provided an initial investigation of expert disagreement in the context of sequential labeling tasks, studying the effectiveness of in-person adjudication for resolving disagreement and for analyzing information about the original source of disagreement.

Our results suggest that majority vote is not necessarily a good proxy for group deliberation decisions in sleep staging. This finding provides some confidence in the usefulness of expert discussions for the purpose of resolving disagreement cases. Beyond that, our qualitative analysis of expert discussion dialogues uncovered a diverse set of different reasons why domain experts disagree in the context of sleep stage classification, most of which go beyond the notion of mere input mistakes.

Perhaps most importantly, we identified two sources of disagreement with a clear connection to the sequential nature of the labeling task and underlying data. This observation provides some support for our hypothesis that the reciprocity of data and context in sequential labeling may lead to unique forms of expert disagreement that are characteristic for sequential labeling tasks, where the interpretation of one case affects the interpretation of subsequent or previous cases. One exciting avenue for future research is the problem of whether it is possible to detect the "critical" tasks that might set up a cascade of disagreement and potentially incorrect labels. Successful detection of such "critical" tasks would allow for a more cost-effective use of expert resources by focusing disambiguation procedures on those cases and saving expert resources on other cases that may be resolved automatically.

Picking up on Penzel et al.'s comment on the nature of "truth" in sleep staging [15], some of the inherent difficulty may arise because there exists a certain degree of both temporal and spatial continuity at transitions between states. In other words, despite the fact that any single neuron or cortical circuit may be thought of as existing in one state or another at any given moment, it is possible for local assemblies of neurons to take some time to transition from one state to another, and also that distant assemblies of neurons in different parts of the brain can exist in different states at the same time. These transitions may take minutes [18, 23] which encompasses several 30-second epochs. Thus, we hypothesize that some of the ambiguity stems from the need to force transitional states into one sleep stage category or another.

We posit that expert disagreement in complex tasks can be used as a signal to identify ambiguous edge cases, and as a driver for eliciting conclusive expert discussions to disambiguate such edge cases. For future work, we propose the idea that example-based training procedures for non-expert crowdworkers may benefit from the presentation of edge cases and their associated expert discus-

sions. While expert disagreement may be one signal for the identification of edge cases, other techniques for the automatic selection of edge case examples, e.g., based on measures of machine uncertainty, have been proposed in prior work [10]. We believe that expert disagreement and the associated expert discussions open up interesting opportunities for optimizing example-based training procedures for human learners, e.g., to improve disambiguation skills and depth of understanding.

Figure 4 illustrates a high-level overview of some of these future directions. In summary, we hope to conduct research on augmenting example-based training procedures for non-expert crowdworkers using edge-cases and their associated expert discussions to help human learners develop more accurate classification strategies for expert-level tasks exhibiting a certain amount of ambiguity.

Another promising avenue for future work will be to explore the minimum "bandwidth" and effective protocols of communication between experts needed to result in successful disambiguation in the context of sequential labeling settings like the one presented in this work. Comparisons may include different styles of expert communication ranging from online text-based asynchronous approaches, to in-person verbal real-time communication.

5 Conclusion

In this work, we reported results from a case study on expert disagreement in sequential labeling tasks where the interpretation of one case can affect the interpretation of subsequent or previous cases. Three board-certified sleep technologists scored 1,920 cases in a sequential 5-class labeling task. Out of all disagreement cases, 30 cases were discussed and resolved through face-to-face adjudication. We identified various sources of disagreement that are specific to the sequential nature of the underlying data and labeling procedure. Our work concluded with a discussion of promising application scenarios of expert discussions for the training of non-expert crowdworkers that we hope to explore in future work.

References

1. Beatty, J., Moore, A.: Should We Aim for Consensus? Episteme **7**(3), 198214 (2010). https://doi.org/10.3366/E1742360010000948
2. Chang, J.C., Amershi, S., Kamar, E.: Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17. pp. 2334–2346. ACM, ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3025453.3026044, http://dl.acm.org/citation.cfm?doid=3025453.3026044
3. Dumitrache, A., Aroyo, L., Welty, C.: Crowdsourcing Ground Truth for Medical Relation Extraction. ACM Transactions on Interactive Intelligent Systems **8**(2), 1–20 (7 2018). https://doi.org/10.1145/3152889, http://dl.acm.org/citation.cfm?doid=3232718.3152889

4. Garbayo, L.: Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. Constraint Programming and Decision Making **539**, 35–45 (2014),
5. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: Modeling individual labelers improves classification. In: AAAI Conference on Artificial Intelligence (2018), https://arxiv.org/pdf/1703.08774.pdf
6. Gurari, D., Grauman, K.: CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17. pp. 3511–3522. ACM, ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3025453.3025781, http://dl.acm.org/citation.cfm?doid=3025453.3025781
7. Jones, A.M.: Victims of Groupthink: A Psychological Study of Foreign Policy Decisions and Fiascoes. The ANNALS of the American Academy of Political and Social Science **407**(1), 179–180 (5 1973). https://doi.org/10.1177/000271627340700115, http://journals.sagepub.com/doi/10.1177/000271627340700115
8. Kairam, S., Heer, J.: Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16. pp. 1635–1646. ACM Press, New York, New York, USA (2016). https://doi.org/10.1145/2818048.2820016, http://dl.acm.org/citation.cfm?doid=2818048.2820016
9. Kiesler, S., Sproull, L.: Group decision making and communication technology. Organizational Behavior and Human Decision Processes **52**(1), 96–123 (6 1992). https://doi.org/10.1016/0749-5978(92)90047-B, http://linkinghub.elsevier.com/retrieve/pii/074959789290047B
10. Kim, J., Park, J., Lee, U.: EcoMeal: A Smart Tray for Promoting Healthy Dietary Habits. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16. pp. 2165–2170. ACM Press, New York, New York, USA (2016). https://doi.org/10.1145/2851581.2892310, http://dl.acm.org/citation.cfm?doid=2851581.2892310
11. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R.: Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology (3 2018). https://doi.org/10.1016/j.ophtha.2018.01.034, http://arxiv.org/abs/1710.01711 http://linkinghub.elsevier.com/retrieve/pii/S0161642017326982
12. Mitry, D., Zutis, K., Dhillon, B., Peto, T., Hayat, S., Khaw, K.T., Morgan, J.E., Moncur, W., Trucco, E., Foster, P.J.: The Accuracy and Reliability of Crowdsource Annotations of Digital Retinal Images. Translational Vision Science & Technology **5**(5), 6 (2016). https://doi.org/10.1167/tvst.5.5.6, http://tvst.arvojournals.org/article.aspx?doi=10.1167/tvst.5.5.6
13. Mumpower, J.L., Stewart, T.R.: Expert Judgement and Expert Disagreement. Thinking & Reasoning **2**(2-3), 191–212 (7 1996). https://doi.org/10.1080/135467896394500, https://www.tandfonline.com/doi/full/10.1080/135467896394500
14. Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., Sigman, M.: Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nature Human Behaviour (1 2018). https://doi.org/10.1038/s41562-017-0273-4, http://www.nature.com/articles/s41562-017-0273-4

15. Penzel, T., Zhang, X., Fietze, I.: Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. Journal of Clinical Sleep Medicine **9**(1), 81–87 (2013)
16. Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y.: Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks (7 2017), http://arxiv.org/abs/1707.01836
17. Rosenberg, R.S., van Hout, S.: The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. Journal of Clinical Sleep Medicine (1 2013). https://doi.org/10.5664/jcsm.2350, http://www.aasmnet.org/jcsm/ViewAbstract.aspx?pid=28772
18. Saper, C.B., Fuller, P.M., Pedersen, N.P., Lu, J., Scammell, T.E.: Sleep State Switching. Neuron **68**(6), 1023–1042 (12 2010). https://doi.org/10.1016/j.neuron.2010.11.032, http://linkinghub.elsevier.com/retrieve/pii/S0896627310009748
19. Schaekermann, M., Goh, J., Larson, K., Law, E.: Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In: Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'18). New York City, NY (2018). https://doi.org/10.1145/3274423
20. Solomon, M.: Groupthink versus The Wisdom of Crowds : The Social Epistemology of Deliberation and Dissent. The Southern Journal of Philosophy **44**(S1), 28–42 (3 2006). https://doi.org/10.1111/j.2041-6962.2006.tb00028.x, http://doi.wiley.com/10.1111/j.2041-6962.2006.tb00028.x
21. Solomon, M.: The social epistemology of NIH consensus conferences. In: Establishing medical reality, pp. 167–177. Springer (2007)
22. Warby, S.C., Wendt, S.L., Welinder, P., Munk, E.G.S., Carrillo, O., Sorensen, H.B.D., Jennum, P., Peppard, P.E., Perona, P., Mignot, E.: Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nature Methods **11**(4), 385–392 (2 2014). https://doi.org/10.1038/nmeth.2855, http://www.nature.com/doifinder/10.1038/nmeth.2855
23. Wright Jr, K.P., Badia, P., Wauquier, A.: Topographical and temporal patterns of brain activity during the transition from wakefulness to sleep. Sleep **18**(10), 880–889 (1995)

Characterising and Mitigating Aggregation-Bias in Crowdsourced Toxicity Annotations

Agathe Balayn[1,2], Panagiotis Mavridis[1], Alessandro Bozzon[1], Benjamin Timmermans[2], and Zoltán Szlávik[2]

[1] TU Delft, Web Information Systems
a.m.a.balayn@student.tudelft.nl,{p.mavridis,a.bozzon}@tudelft.nl
[2] IBM Netherlands, Center for Advanced Studies
{b.timmermans,zoltan.szlavik}@nl.ibm.com

Abstract. Training machine learning (ML) models for natural language processing usually requires large amount of data, often acquired through crowdsourcing. The way this data is collected and aggregated can have an effect on the outputs of the trained model such as ignoring the labels which differ from the majority. In this paper we investigate how label aggregation can bias the ML results towards certain data samples and propose a methodology to highlight and mitigate this bias. Although our work is applicable to any kind of label aggregation for data subject to multiple interpretations, we focus on the effects of the bias introduced by majority voting on toxicity prediction over sentences. Our preliminary results point out that we can mitigate the majority-bias and get increased prediction accuracy for the minority opinions if we take into account the different labels from annotators when training adapted models, rather than rely on the aggregated labels.

Keywords: dataset bias · Machine Learning fairness · crowdsourcing · annotation aggregation.

1 Introduction

When using crowdsourcing to gather training data for Machine Learning (ML) algorithms, several workers work with the same input samples and the annotations are aggregated into a unique one like the majority vote (MV) to ensure its correctness (elimination of annotation mistakes and spammers mainly). Although this data collection method is designed to get high-quality data, we expect that certain tasks involving subjectivity such as image aesthetic prediction, hate speech detection, detection of violent video segments, sentence sentiment analysis, cannot be tackled this way: samples should not be described with unique labels only since they are interpretable differently by different persons.

The use of hate/toxic speech has increased with the growth of the Internet [5]. Predicting whether a sentence is toxic is highly subjective because of its multitude of possible interpretations. The sentence "I agree with that and the fact that the article needs cleaning. Some of these paragraphs [..] seem like they

were written by 5 year olds." is judged negative or positive by different readers, but this perceptions' diversity is ignored when selecting one unique label as done in recent research [4]. [3] studied the existence of identity term biases resulting from the imbalance of a toxicity dataset content, we show with the example of MV-aggregation that crowdsourcing processing methods on the same dataset also create an algorithmic bias here towards the majority opinion. When annotations differ but are all valid for certain annotators, aggregation loses information and leads to decrease of accuracy and unfairness in ML results, thus we hypothesize that the bias can be mitigated by using disaggregated data. In this study, we first exhibit the presence of the majority-bias and its consequences, then we propose a methodology to expose and counter its algorithmic effects.

2 Majority-biased dataset and consequences

We show on the toxicity dataset [6] that in usual crowdsourcing aggregations of annotations, certain worker contributions are ignored for the majority and that it affects the fairness of ML algorithms' results. The dataset consists of 159686 Wikipedia page comments for which 10 annotations per sample are available. A large number of annotators (4301) that we have their personal information rate the phrases with 5 labels of toxicity ranging from -2 (very toxic) to 2 (very healthy) with 0 being neutral.

Subjectivities in the dataset. For each worker, we compute the average disagreement rate (ADR) with the ground truth (percentage of annotations different from the MV here), and plot the distribution over the dataset after removing the annotations of the lowest quality workers (spammers) (fig. 1). The quality score for each worker (WQS) is computed with the CrowdTruth framework [1] using binary labels ([-2;-1]:toxic, [0;2]:non-toxic), along a unit quality score (UQS) to represent the clarity of each sentence. Without removing low-quality workers, the proportion of high agreement is high because most spammers constantly use one positive label and the dataset is unbalanced with more samples with non-toxic MV. The more possible spammers are removed, the more the disagreement increases until the distributions stabilize. Only 0.09% of the workers always agree with the MV for 50 spammers removed: MV-aggregation is not representative of most individuals but only of a sentence-level common opinion.

Algorithmic effect of the bias. We consider the task of predicting binary labels. Training traditional algorithms to predict the MV, annotations of only maximum 0.09% of annotators would be entirely correct: the majority-bias is not consistent with the worker's individual opinions. We evaluate traditional models (sec. 3) trained and tested on aggregated and disaggregated labels (table 1). In both cases accuracy is higher when measured on aggregated data, what shows that classical input data's treatment makes usual models' predictions biased towards one type of opinion, here the majority opinion, instead of representing each subjectivity.

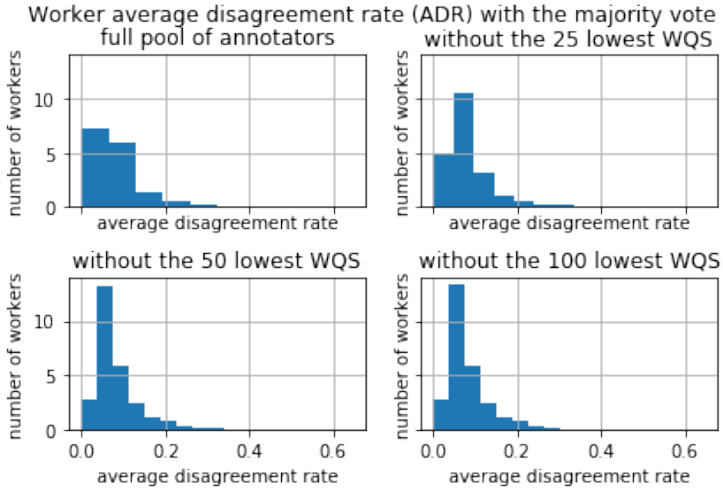


Fig. 1. Normalized distribution comparison of the ADR with the MV with and without low quality worker filtering.

Table 1. Accuracy performances of the model on the ambiguity balanced dataset.

	agg. testing	disagg. testing
agg. training	0.76	0.70
disagg. training	0.77	0.71
disagg. training with user	0.77	0.70

3 Method to measure and mitigate the bias

We claim that a fairer algorithm should return different outputs for a same sample depending on its reader. Here, we propose measures of the majority-bias' algorithmic effect and a method to counter its unfairness.

Bias measure. Global metrics are usually used to optimize the algorithms' parameters and evaluate them. However, they do not inform on the bias' effects since most samples' labels have a high-agreement: the slight improvement when training on disaggregated data hints only lightly at label disaggregation (table 1, fig. 2). To identify the effects, we propose to measure sentence-level and worker-level accuracies on the annotations spread in the following bins: we divide the sentences along their ambiguity score (AS) (percentage of agreement in annotations) or UQS, the workers with their ADR, WQS or demographics categories; and also plot histograms of the per-user and per-sentence errors to identify potential unfairness among all workers or sentences.

Bias mitigation: ML. To account for the full range of valid opinions, we propose to modify the inputs to the ML models. After removing low-quality workers, instead of the aggregated labels we feed them with the annotations augmented with the available worker demographics (age, gender, education, with

a continuous or one-hot encoded representation) that psychology literature [2] gives as the most influencing factors of offensiveness perception (along with ethnicity not available here). Each (sentence, demographics, annotation) tuple is considered as one data sample. We employ the Logistic Regression (LR) classifier, and encode sentences with term frequency-inverse document frequency (tf-idf). The optimal hyperparameters for each set-up are chosen by performing a grid search.

Bias mitigation: dataset balancing. We define 4 data set-ups to help the algorithms learn the individual annotations. Sentence AS and MV-toxicity are computed, and we resample the dataset following the original distribution or balancing the distribution on these 2 criteria, to obtain a dataset whose majority-bias is decreased by equally representing samples with high and low agreement between workers. We also resample the annotations along the MV-toxicity and demographics categories (removing the least frequent ones) into one dataset following the distributions and a balanced one, to foster performance fairness in-between populations.

Results. Binned metrics like the user-level ADR-binned accuracy (fig. 2 with bins along the y-axis) enable to show that models are more suited to workers who agree with the MV (bottom of the y-axis), and highlight the benefit of using disaggregated data with adapted ML models. On the AS-balanced dataset (left part of the x-axis), the user representation increases accuracy for workers with a high disagreement with the majority over using aggregated data or no user-model. The resampling choice also helps understanding and mitigating bias' effects: balancing on demographics neither clearly shows the performance gap between minority and high-ADR workers nor improves accuracy with the user representation, contrary to the AS dataset in which MV-consensus' presence is reduced.

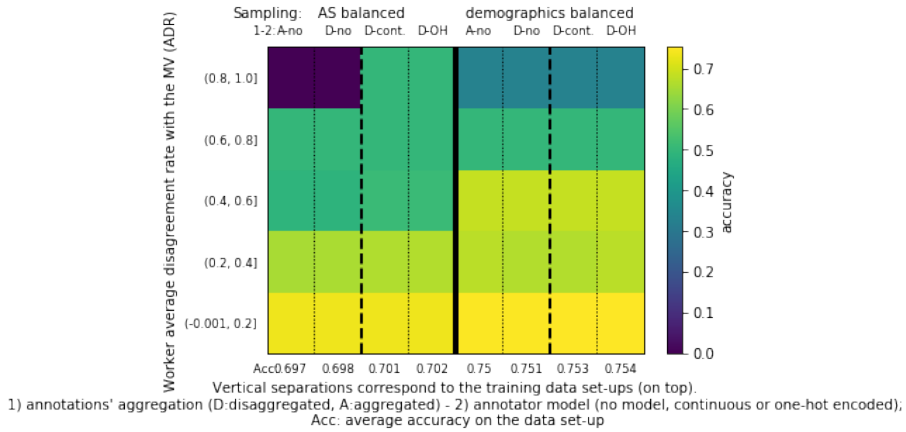


Fig. 2. Average and ADR-binned accuracies for two resamplings of the dataset.

4 Conclusion and Discussion

Disaggregating the annotations decreases the majority-bias' effects with adapted ML models' inputs and dataset resamplings. Binning the evaluation metrics enables to understand and verify the existence of these effects. We only reported results using the LR classifier but we now investigate adaptations of Deep Learning algorithm's architectures which are better suited to the large dataset (10 times more annotations than labels) and to the size of the ML inputs.

Acknowledgements

This research is supported by the Capture Bias project [3], part of the VWData Research Programme funded by the Startimpuls programme of the Dutch National Research Agenda, route "Value Creation through Responsible Access to and use of Big Data" (NWO 400.17.605/4174).

References

1. Aroyo, L., Welty, C.: Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. WebSci2013. ACM **2013** (2013)
2. Cowan, G., Hodge, C.: Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. Journal of Applied Social Psychology **26**(4), 355–374 (1996)
3. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification (2017)
4. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10 (2017)
5. Tsesis, A.: Hate in cyberspace: Regulating hate speech on the internet. San Diego L. Rev. **38**, 817 (2001)
6. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1391–1399. International World Wide Web Conferences Steering Committee (2017)

[3] https://capturebias.eu/

How Biased Is Your NLG Evaluation?

Pavlos Vougiouklis[1†], Eddy Maddalena[2†], Jonathon Hare[1], and Elena Simperl[2]

School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
[1] {pv1e13, jsh2}@ecs.soton.ac.uk
[2] {e.maddalena, e.simperl}@soton.ac.uk

Abstract. Human assessments by either experts or crowdworkers are used extensively for the evaluation of systems employed on a variety of text generative tasks. In this paper, we focus on the human evaluation of textual summaries from knowledge base triple-facts. More specifically, we investigate possible similarities between the evaluation that is performed by experts and crowdworkers. We generate a set of summaries from DBpedia triples using a state-of-the-art neural network architecture. These summaries are evaluated against a set of criteria by both experts and crowdworkers. Our results highlight significant differences between the scores that are provided by the two groups.

Keywords: Natural Language Generation · Human Evaluation · Crowdsourcing

1 Introduction

In the last decade, crowdsourcing has gained increased interest since it offer the methods to reach large amounts of online contributors capable of performing in a small time large amounts of short human intelligence tasks. In particular, it has served the evaluation purposes in different areas of computer science, such as information retrieval [1], machine learning [6], and Natural Language Processing [8].

Human judgements are used for the evaluation of many systems employed on a variety of text generative tasks ranging from Machine Translation [2] and conversational agents [12,13] to generation of summaries [5,3,14] and questions [9,4] in natural language over knowledge graphs. Depending on the task and the evaluation criteria, these judgements are collected by either a small group of "experts" or at a larger scale by crowdworkers that are recruited through a crowdsourcing platform. Especially in the case of Natural Language Generation (NLG) over knowledge graphs, human evaluation is crucial. This is attributed to the inadequacy of the automatic text similarity metrics, such as BLEU [10] or ROUGE [7], to objectively evaluate the generated text [11].

[†]The authors contributed equally to this work.

In this paper, we focus on the human evaluation of textual summaries from knowledge base triple-facts [3,14]. More specifically, we wish to investigate whether there is any similarity between the way that experts and crowdworkers perform on the same evaluation tasks. We compile a list of three criteria that are usually employed for the human evaluation of automatically generated texts [5,14]: (i) fluency, (ii) coverage, and (iii) contradictions. We use the neural network approach that has been recently proposed by Vougiouklis et al. in order to generate textual summaries from DBpedia triples. The summaries are evaluated against the selected criteria by both experts and crowdworkers using the same task interface.

Our experiments have showed that there are significant differences between the scores that are provided by experts and the crowdworkers. Our future work will focus on the methods with which the crowdworkers should be trained in order to perform more accurately on similar tasks.

2 Experimental Design

We run a crowdsourcing task according to which we evaluate 20 summaries that have been generated with the Triples2GRU system that has been proposed by Vougiouklis et al.. We regard each summary as a a concise representation in natural language of an input set of triple-facts. Each summary is generated by Triples2GRU given a set of 8 to 18 triples[1], and is evaluated by 10 workers.

Before starting the task, the workers are presented with general instructions. They are also informed with respect to the ethics approval that we have received for the carrying out of this experiment. The task consists of three phases through which workers were required to evaluate a given summary: (i) text fluency (with an integer number between 1 and 6), (ii) information coverage, by classifying as "Present" or "Absent" each triple-fact from a given list, and (iii) contradictions, by classifying each one of the aforementioned facts as "Direct Contraction" or "Not a Contradiction". At the beginning of each phase, the workers are presented with definitions, suggestions, examples and counter-examples. Each worker was rewarded with 0.20$. After the carrying out of the experiment, the same 20 summaries are also evaluated under the same setup by two experts.

3 Results

Fluency. For each summary, (i) we computed the average of the fluency scores that have been assigned by the 10 workers. Then, (ii) we computed the average of all the values obtained in (i) resulting in an average of 4.8 out of 6. The average fluency with which the experts evaluated the 20 summaries was 5.28. The ANOVA test computed on the two fluency score series produced $p < 0.05$. Consequently, we can claim that compared to the experts, crowdworkers tend to systematically underestimate the summaries' fluency by 0.5 out of 6.

[1]The pre-trained version of Triples2GRU that we used (i.e. https://github.com/pvougiou/Neural-Wikipedian) accepts up to 22 triples as input.

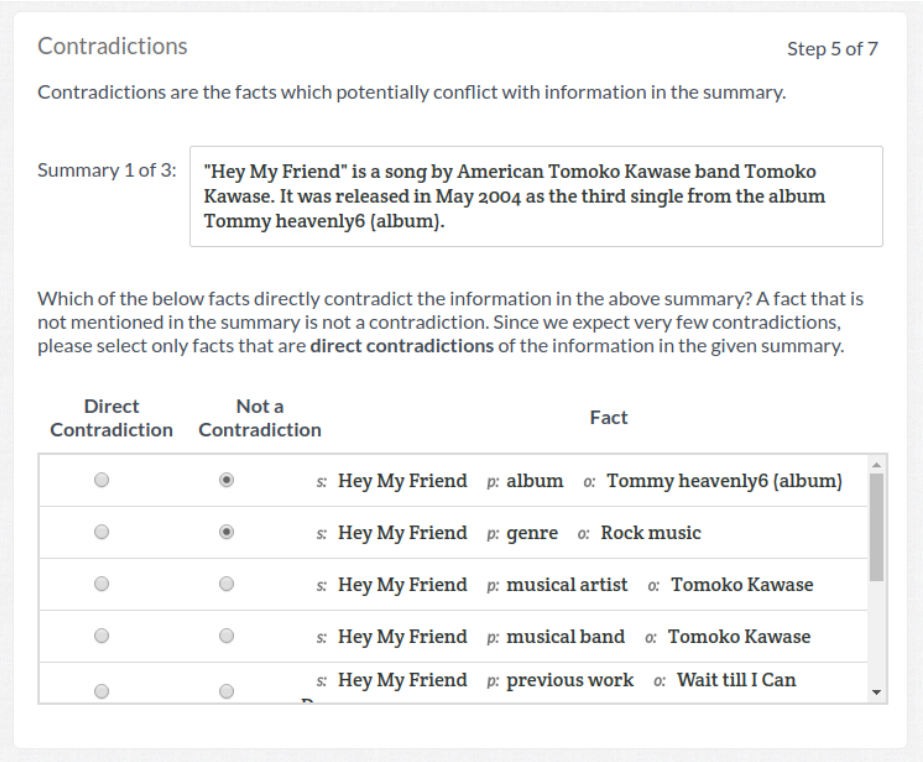


Fig. 1. Task interface showing the page that both the experts and crowdworkers used to identify facts whose information is contradicted in the summary.

Coverage. Workers evaluated the coverage of each summary with respect to a set of triple-facts that generated it. Each summary is aligned with $8 - 18$ facts. The assessments were made by choosing between two labels: (i) "Present" for facts that are either implicitly or explicitly mentioned in the summary, and (ii) "Absent" for the rest. We compute the percentage of the "Present" facts for each summary. Then, similarly to fluency, for each summary, we first compute the average of coverage across the workers, and then the average across all the summaries. The average coverage for all the 20 summaries was 26.85%. In our second experiment, two experts repeated together the same evaluation resulting in an average of 39.71% of facts covered by the summaries. As a result, workers tend to undercount the presence of facts in the generated summaries (confirmed by ANOVA test $p < 0.05$). Finally, a positive significant correlation (Pearson $=$ 0.64) pointed out that workers evaluate coverage in a consistent manner with the experts.

Contradictions. Workers were required to evaluate possible contradictions between the information in a given summary and the respective facts that generated it. Workers were required to mark as "Direct Contradiction" facts that contra-

dict the summary, and as "Not a Contradiction" the rest. For each summary, we compute the percentage of facts that are labelled as contradictions by each single workers. Similarly to coverage, (i) for each summary, we computed the average of the percentages of contradictions of all the workers, and (ii) we averaged the contradiction scores across all the summaries. In a preliminary version of our experiments, each fact was to be marked as either "Contradiction" or "Not Contradiction". However, this proved inadequate since workers were marking facts that were not covered in the summary as contradicting, resulting in an average of $\sim 50\%$ of facts whose information is contradicted in the summaries. In order to minimize the effect of contradictions, besides changing the available labels for each triple-fact, in the contradiction instructions (shown before the third phase of the task), we explicitly noted that contradictions should be rare and that we expected many summaries without any of them. As shown in Fig. 1, we advise workers to identify as contradictions only "Direct contradictions" whose information is explicitly negated in the corresponding summary. Our final result of 30% represents the average of contradicting facts per summary. The same evaluation was made by the two expert, and the average percentage of triple-facts that are contradicted in the summaries was 0.7%. Consequently, workers tend (ANOVA test, $p < 0.05$) to significantly overestimate the presence of facts that are contradicted in the generated summaries.

4 Conclusion

In this paper, we presented preliminary results of a work aimed to explore the use of crowdsourcing for the evaluation of NLG systems. In particular, we focused on the evaluation of textual summaries that are generated from triple-facts. We compared the results of two studies, one that has been performed by experts and one by crowdworkers. The evaluations were conducted in three phases: (i) the fluency of the summary, (ii) the coverage, and (iii) the contradictions of a summary; the latter two are assessed with respect to the given triple-facts. Our preliminary analysis shows that crowdworkers tend to underestimate the fluency of the summaries by 0.5 out of 6. While coverage is judged consistently across both experts and crowdworkers, it is significantly underestimated by the latter. Lastly, despite the fact that we emphasised on the low number of expected contradicting facts, workers strongly overestimated their presence.

A natural extension of this work is to identify the type of facts (i.e. predicates) that influence negatively the workers' judgement. Further studies will focus on minimising this bias by both training workers on how to identify only direct contradictions, and increasing the quality control of the experiment.

Acknowledgements

This research is partially supported by the Answering Questions using Web Data (WDAqua) and QROWD projects, both of which are part of the Horizon 2020 programme under respective grant agreement Nos 642795 and 723088.

References

1. Alonso, O., Mizzaro, S.: Using crowdsourcing for trec relevance assessment. Information processing & management **48**(6), 1053–1066 (2012)
2. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., Turchi, M.: Findings of the 2017 conference on machine translation (wmt17). In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. pp. 169–214. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), http://www.aclweb.org/anthology/W17-4717
3. Chisholm, A., Radford, W., Hachey, B.: Learning to generate one-sentence biographies from Wikidata. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 633–642. Association for Computational Linguistics, Valencia, Spain (April 2017), http://www.aclweb.org/anthology/E17-1060
4. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1342–1352. Association for Computational Linguistics, Vancouver, Canada (July 2017), http://aclweb.org/anthology/P17-1123
5. Ell, B., Harth, A.: A language-independent method for the extraction of RDF verbalization templates. In: Proceedings of the 8th International Natural Language Generation Conference (INLG). pp. 26–34. Association for Computational Linguistics, Philadelphia, Pennsylvania, U.S.A. (June 2014), http://www.aclweb.org/anthology/W14-4405
6. Lease, M.: On quality control and machine learning in crowdsourcing. Human Computation **11**(11) (2011)
7. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Marie-Francine Moens, S.S. (ed.) Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (July 2004)
8. Marujo, L., Gershman, A., Carbonell, J., Frederking, R., Neto, J.P.: Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. arXiv preprint arXiv:1306.4886 (2013)
9. Ngonga Ngomo, A.C., Bühmann, L., Unger, C., Lehmann, J., Gerber, D.: Sorry, i don't speak SPARQL: Translating SPARQL queries into natural language. In: Proceedings of the 22Nd International Conference on World Wide Web. pp. 977–988. WWW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488388.2488473, http://doi.acm.org/10.1145/2488388.2488473
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1073083.1073135
11. Reiter, E.: Natural Language Generation, chap. 20, pp. 574–598. Wiley-Blackwell (2010). https://doi.org/10.1002/9781444324044.ch20, https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444324044.ch20

12. Ritter, A., Cherry, C., Dolan, W.B.: Data-driven response generation in social media. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 583–593. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
13. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B.: A neural network approach to context-sensitive generation of conversational responses. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 196–205. Association for Computational Linguistics, Denver, Colorado (May–June 2015)
14. Vougiouklis, P., ElSahar, H., Kaffee, L., Gravier, C., Laforest, F., Hare, J.S., Simperl, E.: Neural wikipedian: Generating textual summaries from knowledge base triples. CoRR **abs/1711.00155** (2017), `http://arxiv.org/abs/1711.00155`

LimitBias! Measuring Worker Biases in the Crowdsourced Collection of Subjective Judgments

Christoph Hube, Besnik Fetahu and Ujwal Gadiraju

L3S Research Center, Leibniz Universität Hannover
Appelstrasse 4, Hannover 30167, Germany
{hube, fetahu, gadiraju}@L3S.de

Abstract. Crowdsourcing results acquired for tasks that comprise a subjective component (e.g. opinion detection, sentiment analysis) are affected by the inherent bias of the crowd workers. This leads to weaker and noisy ground-truth data. In this work we propose an approach for measuring crowd worker bias. We explore worker bias through the example task of *bias detection* where we compare the worker's opinions with their annotations for specific topics. This is a first important step towards mitigating crowd worker bias in subjective tasks.

1 Introduction

Crowdsourcing is one of the most common means in obtaining ground-truth data for training automated models for a large variety of tasks [9, 7, 6]. In many cases, the annotations are affected by the subjective nature of tasks (e.g. opinion detection, sentiment analysis) or the biases of the workers themselves. For instance, for tasks like determining the political leaning or biased language in a piece of text the annotations, how we perceive something as liberal/conservative or biased is subject to several factors like *framing* and *epistemological* biases in language, social and cultural background of workers, etc. [3].

In this work, we aim at understanding and mitigating the worker biases in crowdsourced annotation tasks that are of subjective nature (e.g., political leaning of a statement, biased language etc.). In particular, we are interested in the case where for a given set of strict annotation rules how does the workers' bias influence their annotation quality. Furthermore, having this setting in mind, how can we mitigate such worker biases in subjective tasks. We propose an approach for measuring crowd worker bias based on the example task of labeling statements as either biased or neutral. In addition to the main task we ask workers for their personal opinion on each statement's topic. This additional information allows us to measure correlations between a worker's opinion and their choice of labeling. In future work we will introduce methods for mitigating the measured bias.

2 Related Work

2.1 Bias in Crowdsourcing Data Acquisition

Recent works have explored task related factors such as complexity and clarity that can influence and arguably bias the nature of task-related outcomes [5]. Work envi-

ronments (i.e., the hardware and software affordances at the disposal of workers) have also shown to influence and bias task related outcomes such as completion time and work quality [4]. In other closely related work, Eickhoff has studied the prevalence of cognitive biases as a source of noise in crowdsourced data curation, annotation and evaluation [1]. Eickhoff studied the effect size of common cognitive biases such as the ambiguity effect, anchoring, bandwagon and decoy effect in a typical relevance judgment task framework. Crowdsourcing tasks are often susceptible to participation biases. This can be further exacerbated by incentive schemes [2]. Other demographic attributes can also become a source of biased judgments. It has also been found that American and Indian workers differed in their perceptions of non-monetary benefits of participation. Indian workers valued self-improvement benefits, whereas American workers valued emotional benefits [8].

In this work, we aim to disentangle the potential sources of worker bias using the example task of *bias detection*. This will be a first holistic approach towards bias management in crowdsourcing.

2.2 The Case of Subjective Annotations

For many tasks such as detecting subjective statements in text (i.e., text pieces reflecting opinions), or biased and framing issues that are often encountered in political discourse [9, 3], the quality of the ground-truth is crucial.

Yano et al. [10] show the impact of crowd worker biases in annotating statements (without their context) where the labels corresponded to the political biases, e.g. *very liberal*, *very conservative*, *no bias*, etc. Their study shows that crowd workers who identify themselves as *moderates* perceive less bias, whereas conservatives perceive more bias in both ends of the spectrum (*very liberal* and *very conservative*). Interestingly, the distribution of workers is heavily biased towards *moderates*. This raises several issues. First, how can we ensure a balanced representation of workers, where for subjective tasks a balanced representation is crucial. Second, which judgments are more reliable having in mind that more conservative workers tend to perceive statements as more biased in both ends of the political spectrum.

In a similar study to, Iyyer et al. [7] showed the impact of the workers in annotating statements with their corresponding political ideology. In nearly 30% of the cases, it was found that workers annotate statements with the presence of a bias, however, without necessarily being clear in the political leaning (e.g. liberal or conservative). While it is difficult to understand the exact factors that influence workers in such cases, possible reasons may be their lack of domain knowledge, respectively the stances with which different political ideologies are represented on a given topic, or it may be the political leanings of the workers themselves. Such aspects remain largely unexplored and given their prevalence they represent significant quality concerns in ground-truth generation through crowdsourcing.

In this work, we aim at addressing these unresolved quality concerns of crowdsourcing for subjective tasks by disentangling all the possible bias factors.

3 Measuring crowd worker Bias

In Section 1 we introduced the problem of measuring crowd worker bias for a crowdsourcing task including a subjective component. In this section we propose an approach for measuring crowd worker bias for the example task of labeling statements as either biased or neutral. The same approach can be used for other tasks as stated in Section 1.

For the example task we use statements from datasets of subjective and opinionated statements that have been extracted from Wikipedia [6] or ideological books [7]. We first create a set of 10 statement groups with each group containing statements for one controversial topic from a list of widely discussed controversial topics, e.g. *abortion*, *capital punishment*, *feminism*. Each statement group contains one main statement that reflects the central pro/against aspect of the controversy, e.g. "Abortion should be legal". In our task design we use the main statement to determine the worker's opinion on the given topic. Furthermore each group contains 4 additional opinionated statements from the dataset that follow the group's topic, two statements that support the main statement and two against it.

To accurately measure worker bias, we divide the task into two subtasks. In the first subtask we show the worker the opinionated statements. The worker has to label each statement as either "biased" or "neutral". We explain the concepts of biased and neutral wording to the workers and give them a guideline when to label a statement as biased or neutral. We give multiple examples for both classes. We additionally provide a third "I don't know" option. The task design for the first subtask is depicted in Figure 1. A similar task design has been used to create a ground truth for the problem of bias detection [6].

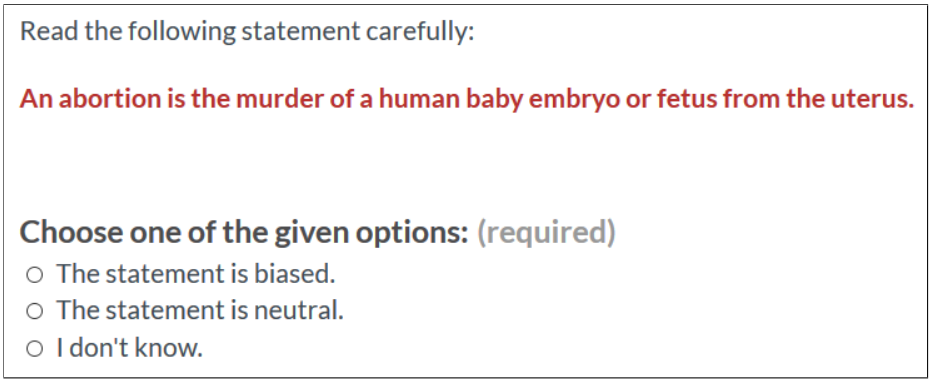


Fig. 1. Main task example.

In the second subtask we ask the worker's opinion for each topic from the statement group. We show the worker the main statement from each group and 5 options on a Likert scale reaching from "I strongly agree" to "I strongly disagree". The task design for the second subtask is depicted in Figure 2.

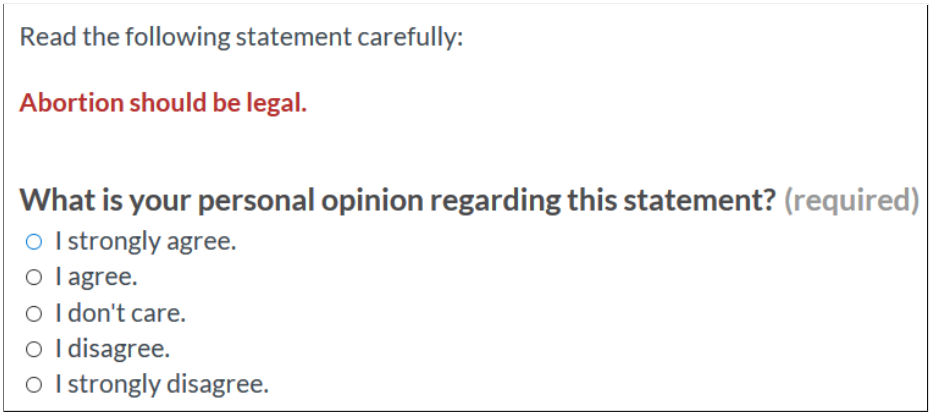


Fig. 2. Opinion example.

Our hypothesis is that workers who agree with a statement are more likely to label it as neutral, i.e. a worker who agrees that abortion should be illegal is more likely to label the statement "An abortion is the murder of a human baby embryo or fetus from the uterus" as neutral. As stated in Section 1 this behavior can negatively influence the crowdsourcing results of this task since crowd workers should label according to the given guidelines and not to personal opinion.

4 Future Work

We introduced an approach for measuring crowd worker bias for crowdsourcing tasks including a subjective component. For future work we are planning to develop methods for mitigating the measured bias. Possible approaches could include balancing judgments between workers of different opinions, making workers aware of their biases (meta-cognition), and discounting strongly biased crowdworkers. Furthermore we want to analyze the influence of task design on worker bias.

Acknowledgments This work is funded by the ERC Advanced Grant ALEXANDRIA (grant no. 339233), DESIR (grant no. 31081), and H2020 AFEL project (grant no. 687916).

References

1. Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 162–170. ACM, 2018.
2. Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.
3. Roger Fowler. *Language in the News: Discourse and Ideology in the Press*. Routledge, 2013.

4. Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):49, 2017.
5. Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 5–14. ACM, 2017.
6. Christoph Hube and Besnik Fetahu. Detecting biased statements in wikipedia. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1779–1786, 2018.
7. Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122, 2014.
8. Ling Jiang, Christian Wagner, and Bonnie Nardi. Not just in it for the money: A qualitative investigation of workers' perceived benefits of micro-task crowdsourcing. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 773–782. IEEE, 2015.
9. Dietram A Scheufele. Framing as a theory of media effects. *Journal of communication*, 49(1):103–122, 1999.
10. Tae Yano, Philip Resnik, and Noah A Smith. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158. Association for Computational Linguistics, 2010.

Investigating Stability and Reliability of Crowdsourcing Output

Rehab Kamal Qarout[1], Alessandro Checco[2], and Kalina Bontcheva[1]

[1] University of Sheffield, Department of Computer Science, Sheffield, UK
[2] University of Sheffield, Information School, Sheffield, UK
{rkqarout1, a.checco, k.bontcheva}@sheffield.ac.uk

Abstract. This research proposes to investigate the reliability of the output of crowdsourcing platforms and its consistency over time. We study the effect of design interface and instructions and identify critical differences between two platforms that have been used widely in research and data collection and evaluation. Our findings will help to uncover data reliability problems and to propose changes in crowdsourcing platforms that can mitigate the inconsistencies of human contributions.

Keywords: crowdsourcing · task design· platforms.

1 Introduction

There are many successful examples on the web of crowdsourcing platforms. However, the features and services provided for the requesters vary from one platform to another, and no single platform meets all the possible requirements that the requesters may have.

We investigate the quality of the output of different platforms when the same task design and dataset is used. To study the reliability and consistency of the output of the platforms and to generalise the findings, we run a continuous evaluation of existing datasets and replicate the task over multiple weeks.

2 Related Work

Crowdsourcing platforms evaluation. In this context, a study by [3] attempts to validate Amazon Mechanical Turk (MTurk) as a tool for collecting data in cognitive behavioural research. They designed several types of experiments and compared the results with traditional laboratory ways of collecting data. The study showed that the quality of the data collected under the experimental conditions in MTurk is highly similar to the quality of the data collected the traditional laboratory way. A similar case study was presented by [1], who analysed the results of surveying the workers on their behaviour of using particular technologies. This research compared the results from MTurk and Survey Monkey to those obtained using a traditional survey. They demonstrated that crowdsourcing platforms can provide the same results and do it much faster when

compared to the traditional way of collecting survey data [1]. Despite some concerns related to the limitations of the technical and visual design of the task and unexpected behaviour such as dropping out of a task before finishing it, collecting data with crowdsourcing saves time and money and reach a wide range of users in a few seconds [3].

A few papers highlighted the differences between crowdsourcing platforms. In one of the recent studies, [6] introduced the new platform Prolific Academic (ProA) and compared the result of this platform with CrowdFlower (CF) and MTurk. The findings of this study recorded the highest response rate for participants in CF and the highest data quality for the participants in ProA and comparable to MTurk's [6]. Another study [5] used Rankings website to collect data and compare crowdsourcing platforms over two periods of time and according to a number of criteria: *type of service provided, quality and reliability, region, online imprint*. The findings of this study discuss the effect of the platforms characteristics of their traffic data and popularity [5, 4].

Time consistency of tasks. studies by [2] investigate the creation of evaluation campaigns for the semantic search task of keyword-based ad-hoc object retrieval using crowdsourcing task. They used a sample of entity-queries from the Yahoo! log and Microsoft log to evaluate the semantic search result. They prove that the reliability of crowdsourcing workers and the quality of the result was comparable to that of the experts even when repeating the same task over time [2]. Following this work, [8] extend the continuous evaluation of information retrieval (IR) systems using crowdsourced relevance judgments.

3 Research Questions

This research will address the following questions:

- **RQ1:** *Is there a significant difference in the quality, reliability, and consistency of the results for the same task repeated over a different time scale?*
- **RQ2:** *Is there a significant difference in the quality, reliability, and consistency of the results for the same task performed on different platforms?*

Answering RQ1 requires conducting a study where the same experiments will be repeated on a different time scale. We replicated the experiment using the same part of the dataset for the same assumption discussed in [2, 7] for measuring repeatable and reliable evaluation over crowdsourcing systems. These studies show experimental proofs that a crowdsourcing platform produces a scalable and reliable result over a repetition time of one month. We examined the consistency of the same task over a shorter time scale (once a week).

RQ2 offers an in-depth analysis and practical comparison of crowdsourcing platforms. We investigated the replication of the same task over multiple crowdsourcing platforms and over different levels of workers' experience and accuracy as provided by each platform. Two of the most popular platforms, that have

been used in crowdsourcing business and research studies of data evaluation and acquisitions, that is, Amazon Mechanical Turk (MTurk) and Figure Eight (F8), have been chosen for this study.

For both research questions and for each platform, we ran multiple types of tasks and measured the stability of the performance over the variations of the following factors:

- The quality of the task interface.
- The workers' experience level provided by the platform.

The evaluation of these factors depended on the completion time of the task and accuracy of the result. Moreover, with repeating the same task every week, the overall time of completing the batch on each platform will be recorded.

4 Experimental Results: Phase 1

The experiments in this phase used the plain interface similar to the one presented in [2]. We repeated the same experiment five times (once every week) and it was launched on the same day of the week and at the same time on each platform. Each task consisted of 20 tweets to be judged by 150 workers. The workers were rewarded with 0.15$ and they could do the task only once since after they finished they were excluded from participating in another batch of the task.

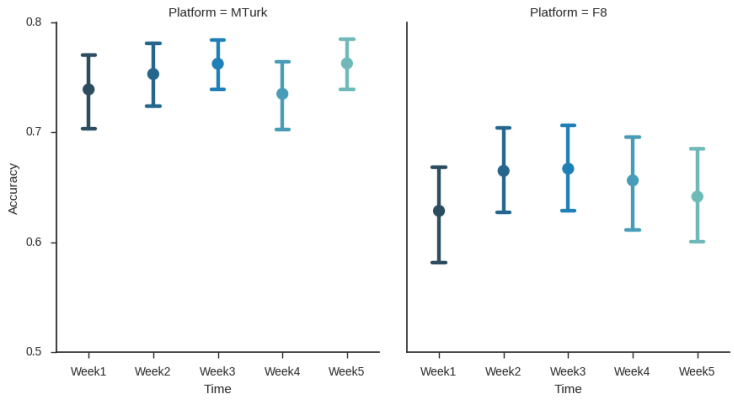


Fig. 1. Accuracy distribution over time.

Table 1 presents the results of the baseline phase experiments for the tweets dataset with comparison between the two selected platforms. The results show some consistency over the five runs on each platform. Workers were finishing the task faster in MTurk, where the average time per assignment was approximately 4 minutes, while it took approximately 6 minutes in F8. The overall accuracy

for each run on MTurk was more than 73% whereas it was in the range of 60% on F8 as shown in Figure 1. Although the results from MTurk are significantly better than those from F8, the total completion time for the whole batch took an average of 3 days in MTurk and 4 to 7 hours in F8.

Table 1. Results of five runs in MTurk and F8

	MTurk	F8
Average Time per Assignment	4 m, 16 s	6 m, 09 s
	4 m, 49 s	6 m, 33 s
	4 m, 24 s	6 m, 18 s
	4 m, 25 s	5 m, 30 s
	4 m, 37 s	5 m, 49 s
Avg.Accuracy & Standard deviation	0.73 ± 0.20	0.63 ± 0.28
	0.76 ± 017	0.66 ± 0.25
	0.76 ± 0.14	0.67 ± 0.25
	0.74 ± 0.19	0.66 ± 0.27
	0.76 ± 0.14	0.64 ± 0.28
Completion Time for the Batch	3 d, 00 h, 14 m	05 h, 11 m
	3 d, 01 h, 29 m	04 h, 45 m
	2 d, 08 h, 36 m	07 h, 10 m
	3 d, 13 h, 54 m	04 h, 43 m
	3 d, 03 h, 28 m	04 h, 04 m

A two-way ANOVA was conducted to examine the effect of repeating the same task several times and on two different platforms on the accuracy of the results (Table 2). There was a statistically significant interaction between the effects of repeating the task on different platforms on the accuracy $p < 0.05$. There were no differences between running the experiment several times on each platform which indicates the consistency of the outcome of each platform. We will investigate the reasons for having this difference accuracies.

Table 2. Results of 2 ways ANOVA test.

	sum_sq	df	F	PR(>)
C(Platform)	3.65	1.0	71.4	0.68e-16
C(Time)	0.17	4.0	0.84	0.49
C(Platform):C(Time)	0.08	4.0	0.42	0.80
Residual	76.17	1490.0	NaN	NaN

5 Future Directions

For the advanced phases of this study, we will investigate why we had these results in the first phase. One of the reasons could be the workers' diversity and

their level of experience. Another reason could be the variation of the amount of payment for different channels in F8. With this study, we hope to reach a reasonable level of understanding what are the best strategies and advise crowdsourcing users on the best way to achieve better service from the system.

6 Acknowledgements

This project is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732328.

References

1. Bentley, F.R., Daskalova, N., White, B.: Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17. pp. 1092–1099 (2017). https://doi.org/10.1145/3027063.3053335
2. Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S.: Repeatable and Reliable Search System Evaluation using Crowdsourcing. Journal of Web Semantics **21**, 923–932 (2011). https://doi.org/10.1016/j.websem.2013.05.005
3. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE **8**(3) (2013). https://doi.org/10.1371/journal.pone.0057410
4. Mourelatos, E., Frarakis, N., Tzagarakis, M.: A Study on the Evolution of Crowdsourcing Websites. ISSNOnline) European Journal of Social Sciences Education and Research **11**(1), 2411–9563 (2017)
5. Mourelatos, E., Tzagarakis, M., Dimara, E.: A REVIEW OF ONLINE CROWDSOURCING PLATFORMS. South-Eastern Europe Journal of Economics **14**(1), 59–74 (2016)
6. Peer, E., Samat, S., Brandimarte, L., Acquisti, A.: Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology **70**(January), 153–163 (2016). https://doi.org/10.1016/j.jesp.2017.01.006
7. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Combining inverted indices and structured search for ad-hoc object retrieval. In: SIGIR. p. 125 (2012). https://doi.org/10.1145/2348283.2348304
8. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Pooling-based continuous evaluation of information retrieval systems. Information Retrieval **18**(5), 445–472 (2015). https://doi.org/10.1007/s10791-015-9266-y

A Human in the Loop Approach to Capture Bias and Support Media Scientists in News Video Analysis

Panagiotis Mavridis[1], Markus de Jong[2], Lora Aroyo[3], Alessandro Bozzon[1], Jesse de Vos[5], Johan Oomen[5], Antoaneta Dimitrova[3], and Alec Badenoch[4]

[1] TU Delft, Web Information Systems
{p.mavridis,a.bozzon}@tudelft.nl
[2] Vrije Universiteit Amsterdam, User-Centric Data Science Group
{lora.aroyo,m.a.dejong}@vu.nl
[3] Leiden University
a.l.dimitrova@fgga.leidenuniv.nl
[4] Utrecht University
A.W.Badenoch@uu.nl
[5] Beel en Geluid
{joomen,jdvos}@beeldengeluid.nl

Abstract. Bias is inevitable and inherent in any form of communication. News often appear biased to citizens with different political orientations, and understood differently by news media scholars and the broader public. In this paper we advocate the need for accurate methods for bias identification in video news item, to enable rich analytics capabilities in order to assist humanities media scholars and social political scientists. We propose to analyze biases that are typical in video news (including framing, gender and racial biases) by means of a human-in-the-loop approach that combines text and image analysis with human computation techniques.

Keywords: Bias detection · bias in news video files · machine learning · crowdsourcing · human computation · human in the loop

1 Introduction

News media scholars analyze online media for different international events from a variety of online news channel sources such as CNN, France24, RT or Al Jazeera. However, news reporters in each channel present news stories from different perspectives. As such, news often appear biased to citizens with different political orientations and are understood differently by news media scholars and broader public [7]. Since bias is inherent in every communication, it could lead to a misguided audience, whether scientists or broader public. For instance it could affect democratic institutions by affecting voters choice [11, 6]. A more accurate detection of bias could enable consumers of video news item to become aware of

possible misrepresentations, and could provide with more useful media analysis for scientists.

Since news media are abundant and manual detection of bias is costly – both in monetary and temporal terms – we propose to assist media news scholars with automatic techniques. Focusing either on the different manifestations of bias or on the ambiguity of interpretations of media news, this problem can be studied from two different perspectives: (1) the study of the different manifestations of bias; and (2) the role of content ambiguity in the detection of bias. In this work, we propose an approach for the first.

2 Related Work

Bias is often manifested through misrepresentation of entities which is performed by framing [1, 12]. Framing is also used when news agencies adjust their report approach for their intended public and target specific groups [3]. The framing acts upon concepts or entities of the story; when such entities are individuals, bias can manifest in terms of (1) gender bias [8] and (2) racial bias [5] when a particular gender or race is misrepresented.

Framing can be captured through either an extensive manual thematic analysis [12] or by word-based quantitative text analysis performed manually or with computer assisted methods [1]. In the case of video, crowdsourced labels have been used to gain insight into how exactly themes and sentiment differ between news sources [3]. As mentioned, research can discover racial bias expressed by discrepancies between on-screen representation of ethnic groups and various official statistics [5]. Example results from this 2017 crowdsourced investigation in Los Angeles showed, for example, that whites were significantly overrepresented in the victim, perpetrator and police officer categories. Similar quantitative comparisons can be carried out to investigate gender bias [8]

However, automated methods for the detection of bias also exist. For instance [9] identify on a particular controversial topic (Edward Snowden) two different groups of twitter users talk about the controversial topic and how information is shaped and propagated about the topic by comparing the rate of original tweets and retweets over this controversial topic during a month. On a similar subject but with a different method, [10] identifies seed words and trains a semi-automatic method to detect partisans on a controversial topic. [4] identifies unintended bias that comes from an imbalanced dataset when demographics on participants are not always available.

3 Proposed Approach

To address bias in news video, we propose a comparative correlation and sentiment analysis of the different manifestations of bias as mentioned in Section 2 through the use case of news analysis for media scientists. We propose to automate a procedure that extracts different properties and elements that can lead to automatic bias detection and involves humans in the loop in an iterative process.

Since the automatic methods are not enough to identify the bias cues related to entities and sentiments we automate a process that involves humans in the loop. Then, social science and political science scholars evaluate the output of this process. More specifically, we specify the initial datasets and explain the preprocessing of the data in order to extract the different bias cues for framing, gender and racial bias with the use of machine learning and human computation methods. In the end we evaluate with the help of our experts.

3.1 Datasets

Videos and textual data: The datasets consist of online news videos reporting on a news event. We gather video and their metadata such as subtitles, video comments and video tags. As sources, we have selected English language online video news channels that post their videos on YouTube and are mentioned in Section 1 as these present international news from different perspectives. We also take advantage of the keyword annotated datasets on videos provided in the YouTube8m dataset[6].

To determine news events we use Wikipedia [7] and online news articles. Wikipedia provides crowd-sourced articles from different contributors. This data takes some time to build, improves over time and could be used to compare the entities and facts presented between different news sources. Online news articles can provide with comparison data over videos when Wikipedia articles are missing.

3.2 Data preprocessing

Captions and Text Extraction: Since we want to compare the video event coverage with online news articles that contain mainly text, we need to retrieve the text mentioned in the video. Thus, we need to generate subtitles for the videos (if none available) using a speech to text engine. We also detect and extract informative text displayed on screen as part of the narration *.e.g.* speaker or location descriptors, section titles) using optical character recognition (OCR).

News event detection and data gathering: From the Wikipedia pages, we extract events using NLP processing. From these events and supported by Wordnet [8] we can create seed words to assist a crowd to annotate an event. When the events are identified, we can collect video data from the different video channels of our initial dataset.

3.3 Bias Cues extraction

We identify the different bias cues by a comparative analysis of the different textual and video data that we have from different sources concerning the same

[6] https://research.google.com/youtube8m/
[7] www.wikipedia.com
[8] wordnet.princeton.edu/

event. This method permits identifying missing or misrepresented entities in terms of number or sentiment attached and thus provides a detection of framing and misrepresentation of gender or race within the presented video. For instance, how many times some entities appear more compared to the other entities on a particular event. We perform the above with different ways such as video deconstruction, keyword and entity extraction and sentiment analysis.

Video deconstruction and Analysis. In order to be able to annotate videos for their events we need to be able to separate the scenes from each video with automated scene recognition. We plan to obtain bias cues with both machine learning and human computation. Ideally, we use machine learning to identify what needs to be annotated by humans in order to find out *e.g.* who is reporting, who is talking, who is present at the scene, etc.

Entity and Sentiment Analysis. To make use of all data modalities in our news videos, we investigate the combination of existing API's for textual, voice- and face-based sentiment analysis [13] attached to entities. Also to be able to attach the entities to particular sentiments [2] we use human computation to identify or validate the output from sentiment analysis from machine learning methods.

3.4 Evaluation

Finally, we evaluate our approach with domain experts from humanities and political sciences. Given an event, they are presented with an interface with different graphs from our hybrid human-machine approach. The expert should be able to use a representation of the event and different word clouds for the same event from different channels and be able to perform the bias investigation.

4 Discussion and Directions

We presented how bias is manifested and can be captured with an approach using state of the art machine learning and human computation. We mainly focused on identifying the different bias cues such as framing, gender and race misrepresentations in order to assist media scientists in video news analysis. We want to apply this approach through a pilot experiment and compare the different types of bias, their possible correlations and also perform a sentiment analysis.

Acknowledgements

This research is supported by the Capture Bias project [9], part of the VWData Research Programme funded by the Startimpuls programme of the Dutch National Research Agenda, route "Value Creation through Responsible Access to and use of Big Data" (NWO 400.17.605/4174).

[9] https://capturebias.eu/

References

1. Boräng, F., Eising, R., Klüver, H., Mahoney, C., Naurin, D., Rasch, D., Rozbicka, P.: Identifying frames: A comparison of research methods. Interest Groups & Advocacy **3**(2), 188–201 (2014)
2. Calais Guerra, P.H., Veloso, A., Meira, Jr., W., Almeida, V.: From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 150–158. KDD '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2020408.2020438, http://doi.acm.org/10.1145/2020408.2020438
3. Dimitrova, A., Frear, M., Mazepus, H., Toshkov, D., Boroda, M., Chulitskaya, T., Grytsenko, O., Munteanu, I., Parvan, T., Ramasheuskaya, I.: The elements of russias soft power: Channels, tools, and actors promoting russian influence in the eastern partnership countries (2017)
4. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification (2018)
5. Dixon, T.L.: Good guys are still always in white? positive change and continued misrepresentation of race and crime on local television news. Communication Research **44**(6), 775–792 (2017)
6. Gelman, A., Azari, J.: 19 things we learned from the 2016 election. Statistics and Public Policy **4**(1), 1–10 (2017). https://doi.org/10.1080/2330443X.2017.1356775, https://doi.org/10.1080/2330443X.2017.1356775
7. Hackett, R.A.: Decline of a paradigm? bias and objectivity in news media studies. Critical Studies in Mass Communication **1**(3), 229–259 (1984). https://doi.org/10.1080/15295038409360036, https://doi.org/10.1080/15295038409360036
8. Kinnick, K.N.: Gender bias in newspaper profiles of 1996 olympic athletes: A content analysis of five major dailies. Women's Studies in Communication **21**(2), 212–237 (1998)
9. Liao, Q.V., Fu, W.T., Strohmaier, M.: #snowden: Understanding biases introduced by behavioral differences of opinion groups on social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 3352–3363. CHI '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858422, http://doi.acm.org/10.1145/2858036.2858422
10. Lu, H., Caverlee, J., Niu, W.: Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 213–222. CIKM '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806573, http://doi.acm.org/10.1145/2806416.2806573
11. N., D.J., Michael, P.: The impact of media bias: How editorial slant affects voters. Journal of Politics **67**(4), 1030–1049. https://doi.org/10.1111/j.1468-2508.2005.00349.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2508.2005.00349.x
12. Philo, G., Briant, E., Donald, P.: Bad news for refugees. Pluto Press (2018)
13. Poria, S., Peng, H., Hussain, A., Howard, N., Cambria, E.: Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing **261**, 217–230 (2017)

Device-Type Influence in Crowd-based Natural Language Translation Tasks

Michael Barz[1], Neslihan Büyükdemircioglu[2], Rikhu Prasad Surya[2], Tim Polzehl[2], and Daniel Sonntag[1]

[1] German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Saarbrücken, Germany
{michael.barz,daniel.sonntag}@dfki.de
[2] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
neslihan.bueyuekdemircioglu@tu-berlin.de,
{rikhu.p.s,tim.polzehl}@qu.tu-berlin.de

Abstract. The effect of users' interaction devices and their platform (mobile vs. desktop) should be taken into account when evaluating the performance of translation tasks in crowdsourcing contexts. We investigate the influence of the device type and platform in a crowd-based translation workflow. We implement a crowd translation workflow and use it for translating a subset of the IWSLT parallel corpus from English to Arabic. In addition, we consider machine translations from a state-of-the-art machine translation system which can be used as translation candidates in a human computation workflow. The results of our experiment suggest that users with a mobile device judge translations systematically lower than users with a desktop device, when assessing the quality of machine translations. The perceived quality of shorter sentences is generally higher than the perceived quality of longer sentences.

Keywords: Crowd-based Translation · Natural Language Translation · Machine Translation · Human Judgment · Crowdsourcing.

1 Introduction

Nowadays, crowdsourcing is used for a variety of tasks ranging from image tagging to text creation and translation [2, 10, 7]. Incorporating humans in complex workflows introduces several challenges including a large variety in their contribution quality [5]. Recent research investigates approaches in which humans are included, if a machine learning model is uncertain, for example, in the domain of natural language translation.

We consider crowd-enabled natural language translation, particularly workflows in which human translators compete against machine translation systems that are developed for low cost and high-speed [1]. Previous research has shown that crowdsourced translations are of higher quality than machine translations, but professional human translators still outperform the crowd [8, 6]. Hence, rollouts of respective business applications fail due to a lack quality in automated

translation and require a human quality assurance. Several concepts and workflows are proposed for ensuring high translation quality, e.g., Minder and Bernstein [8] investigate the suitability of iterative and parallel workflow patterns for generating translations of high quality. Zaidan et al. [11] propose a model for automatically selecting the best translation from multiple translation candidates and calibrate it using professional reference translations. In the domain of machine translation, common metrics for quality assessment include human judgements, but also automated measures that compare translation candidates against reference translations [3]. Gadiraju et al. [4] investigate the effect of the device type on the quality of different crowd tasks, but did not include translation.

In this work, we focus on the influence of the device type of the human assessor on its quality assessment in a crowd-based translation setting and for machine translation. We present our preliminary results of a corresponding experiment in which the crowd was asked to translate and rate a subset of the IWSLT parallel corpus[3]. In addition, we asked them to rate machine translations of the same sentences.

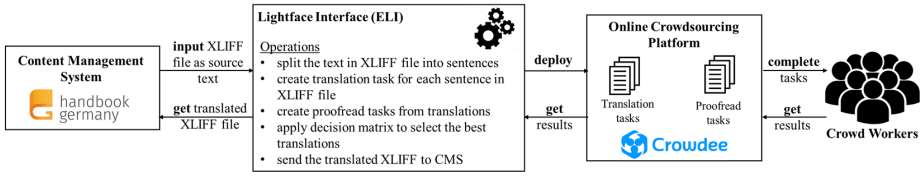


Fig. 1. Workflow diagram of the considered crowd-based translation system.

2 Crowd-Translation System

We implement a simple crowd-based workflow to investigate biases in the quality assessment of crowd-based translations. Our system is implemented using the crowdsourcing platform Crowdee [9], as it has shown to meet scientific requirements in the past, and seamlessly integrates into the enterprise-level content management system (CMS) Adobe Experience Manager, which can be used for administrating the content of multilingual websites, e.g., the refugee information portal handbookgermany.de (see figure 1). Our prototype consists of a combination of iterative and parallel processes including a translation and a proofreading/assessment task to obtain translations of adequate quality. In this setting, we investigate the behavior of human judgments depending on the device type used for the assessment task. As machine translations are commonly used for generating translation candidates, we investigate the same for translations of

[3] https://sites.google.com/site/iwsltevaluation2016/mt-track

the state-of-the-art machine translator Google translate[4]. For a given article, our workflow generates sentence-based translation tasks which can be processed in parallel. Resulting translations are used to create proofread tasks, which ask the crowd to rate and, if necessary, improve the candidate. We use these ratings for our evaluation. One aim of this work is to find suitable metrics based on human judgments and incorporating the inherent bias for (semi-)automatically selecting the best translations.

3 Experiment

For our experiment, we use a subset of the parallel IWSLT evaluation corpus including English transcriptions of TED talks and reference translations in Arabic. We selected an article focusing on climate change[5]. We recruit bilingual crowdworkers via social media channels targeting countries where most people speak English or Arabic. We ask these crowdworkers to participate in language proficiency tests for both languages designed by native speakers. Workers that reach a proficiency of 80% or higher in both tests are selected for participation. For the translation stage, we collect 3 translations for each sentence resulting in a total of 180 translation tasks. Subsequently, we publish 3 proofread tasks for each candidate yielding about 540 human judgments on 5-pt Likert scales. Please note, the actual number of analyzed judgements differs due to illegal or rejected crowd contributions and parallel execution of task repetitions. Overall, we limit the maximum number of translation tasks per crowdworker to 3, in order to include more workers. Similarly, we ask crowd workers to rate machine translations of the source sentences. The human judgment constitutes the dependent variable, the device type used for the assessment task is the independent variable in our experiment. Further, we observe the sentence length as a control variable, as we expect longer sentences to achieve lower quality judgments due to, e.g., lower translation quality or lower perceived quality. We consider two device types, mobile and desktop devices, and split the sentence length into a low and high group based on the median length: we split at 12.5. We apply Kruskal-Wallis tests for significant differences between groups on a 1% significance level, a standard procedure for an analysis of variance for non-parametric distributions and robust against unequally sized groups.

4 Results & Discussion

Concerning human judgments for crowd-translations ($n = 662$), we do not see a significant difference between quality assessments from mobile and desktop users. As a potentially influencing factor, we have only 83 samples from mobile users, yielding a very unbalanced dataset in contrast to our data concerning the machine translations. However, we do observe that human quality judgments are

[4] generated using https://cloud.google.com/ml-engine/
[5] TED talk with ID 535 from TED2009; segments 1 to 60.

significantly lower for long sentences ($Mdn = 4.16$) compared to short sentences ($Mdn = 4.33$). The overall human judgment is $Mdn = 4.3$ ($SD = .69$), which can be interpeted as good overall translation quality.

Concerning the human judgments for machine translations ($n = 163$), we observe that quality assessments from users with mobile devices ($Mdn = 3.55, n = 75$) are lower than those submitted with a desktop device ($Mdn = 3.93, n = 88$). For mobile users, this includes 41 assessments for short sentences and 33 assessments for longer ones. We observed a similar ratio for desktop users: 50 for short and 38 for long sentences. Further, we observe that long and short sentences have approximately the same frequency in the mobile and in the desktop group. This supports the implication that the differences in the quality assessments are induced by the device type and not by an unbalanced distribution of long and short sentences in each group. These findings are in line with the findings of Gadiraju et al. [4]: Using mobile devices negatively impacts the result of crowdtasks. Here, a lower usability might be the cause for systematically lower quality assessments. However, additional factors originating from the workflow design might as well influence the quality assessments which are not taken into account in this paper.

5 Conclusion

We investigated the bias introduced by the device type used for assessing translation quality in crowd-based translation workflows. The results of our study suggest that we can confirm our hypothesis that users assessing translations with the mobile phone provide systematically lower results. This should be taken into account for, e.g., automated translation candidate selection based on human judgments. However, we reject generalizing this statement due to small amount of data included here. Future work should investigate this aspect on a more complete dataset; it should also include further factors that might add a bias to the quality assessment. Ongoing work includes language proficiency and user characteristics. In additon, we found a decline in translation quality for different length of sentences, which is subject to ongoing work on analysis whether this originates from actually lower translation performance on longer sentences or whether is is rather due to a higher task complexity.

6 Acknowledgments

We want to thank EIT Digital for supporting our research project ERICS and our collaborators in this project from T-Systems MMS, Aalto University and Crowdee.

References

1. Barz, M., Polzehl, T., Sonntag, D.: Towards hybrid human-machine translation services. EasyChair Preprint no. 333 (EasyChair, 2018). https://doi.org/10.29007/kw5h

2. Borromeo, R.M., Laurent, T., Toyama, M., Alsayasneh, M., Amer-Yahia, S., Leroy, V.: Deployment strategies for crowdsourcing text creation. Information Systems **71**, 103–110 (2017)
3. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., Way, A.: Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics **108**(1), 109–120 (jan 2017). https://doi.org/10.1515/pralin-2017-0013, http://www.degruyter.com/view/j/pralin.2017.108.issue-1/pralin-2017-0013/pralin-2017-0013.xml
4. Gadiraju, U., Checco, A., Gupta, N., Demartini, G.: Modus operandi of crowd workers: The invisible role of microtask work environments. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(3), 49 (2017)
5. Goto, S., Ishida, T., Lin, D.: Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes. In: Fourth AAAI Conference on Human Computation and Crowdsourcing (2016)
6. Hu, C., Bederson, B.B., Resnik, P., Kronrod, Y.: MonoTrans2 : A New Human Computation System to Support Monolingual Translation. Chi '11 pp. 1133–1136 (2011). https://doi.org/10.1145/1978942.1979111
7. Malone, T.W., Rockart, J.F.: Computers, networks and the corporation. Scientific American **265**(3), 128–137 (1991)
8. Minder, P., Bernstein, A.: How to translate a book within an hour: towards general purpose programmable human computers with crowdlang. In: Proceedings of the 4th Annual ACM Web Science Conference. pp. 209–212. ACM (2012)
9. Naderi, B., Polzehl, T., Wechsung, I., Köster, F., Möller, S.: Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
10. Ross, J., Irani, L., Silberman, M., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: CHI'10 extended abstracts on Human factors in computing systems. pp. 2863–2872. ACM (2010)
11. Zaidan, O.F., Callison-Burch, C.: 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. pp. 1220–1229. Association for Computational Linguistics (2011)