Journal of Artificial Intelligence Research 72 (2021) 1385-1470 Submitted 02/2021; published 12/2021

Learning from Disagreement: A Survey

Alexandra N. Uma A.N.UMA@QMUL.AC.UK
Queen Mary University of London

Tommaso Fornaciari FORNACIARI.TOMMASO@UNIBOCCONI.IT
Dirk Hovy DIRK.HOVY@UNIBOCCONI.IT
Università Bocconi, Milano

Silviu Paun S.PAUN@QMUL.AC.UK
Queen Mary University of London

Barbara Plank BAPL@ITU.DK
IT University of Copenhagen

Massimo Poesio M.POESIO@QMUL.AC.UK
Queen Mary University of London

Abstract

Many tasks in Natural Language Processing (NLP) and Computer Vision (CV) offer evidence that humans disagree, from objective tasks such as part-of-speech tagging to more subjective tasks such as classifying an image or deciding whether a proposition follows from certain premises. While most learning in artificial intelligence (AI) still relies on the assumption that a single (gold) interpretation exists for each item, a growing body of research aims to develop learning methods that do not rely on this assumption. In this survey, we review the evidence for disagreements on NLP and CV tasks, focusing on tasks for which substantial datasets containing this information have been created. We discuss the most popular approaches to training models from datasets containing multiple judgments potentially in disagreement. We systematically compare these different approaches by training them with each of the available datasets, considering several ways to evaluate the resulting models. Finally, we discuss the results in depth, focusing on four key research questions, and assess how the type of evaluation and the characteristics of a dataset determine the answers to these questions. Our results suggest, first of all, that even if we abandon the assumption of a gold standard, it is still essential to reach a consensus on how to evaluate models. This is because the relative performance of the various training methods is critically affected by the chosen form of evaluation. Secondly, we observed a strong dataset effect. With substantial datasets, providing many judgments by high-quality coders for each item, training directly with soft labels achieved better results than training from aggregated or even gold labels. This result holds for both hard and soft evaluation. But when the above conditions do not hold, leveraging both gold and soft labels generally achieved the best results in the hard evaluation. All datasets and models employed in this paper are freely available as supplementary materials.

1. Introduction

Modern research in cognitive science and artificial intelligence (AI) is driven by the availability of large datasets annotated with human judgments (Ide and Pustejovsky, 2017). These data instances and their corresponding labels are not only used to train and test computational models, but also to provide data-driven evidence of linguistic phenomena, complementing a linguist's intuition. In addition, they can be used to compute statistics about the frequencies of certain phenomena (de Marneffe and Potts, 2017).

The simplest way to create an annotated dataset is to appoint a single expert, motivated either by altruism or a financial incentive, to provide the labels for all data instances (or items). But this approach is only feasible for the small-to-medium scale annotations that were the norm until about ten years ago, not for the much larger datasets required today. In addition, the quality of the data produced this way is overly dependent on the expertise of this sole annotator and their skillfulness at annotation. Furthermore, in tasks with elements of subjectivity or ambiguity, the data will be implicitly encoded with any bias the annotator may have about the subject matter. To mitigate these limitations, most large-scale annotation projects use several experts for annotation. Typically, these experts provide 2–3 annotations for each item. A subsequent **adjudication** step produces a single label for each item, the so-called **gold label**. This strategy has been used to annotate most large NLP corpora, for example ONTONOTES (Pradhan et al., 2011; Hovy et al., 2006). However, using experts is very expensive, prohibitively so for large-scale projects. Thus, a third alternative has gained increasing popularity: sourcing annotations from a "crowd" of people, typically (but not always) non-experts. This approach is called **crowdsourcing** (Snow et al., 2008; Michelucci, 2013; Poesio et al., 2017). Crowd workers are usually recruited by offering small financial incentives (in which case the approach is sometimes known as **microtask crowdsourcing**) or by re-configuring the task as a game people play willingly (so-called **game-with-a-purpose** (von Ahn and Dabbish, 2008; Lafourcade et al., 2015)). Crowdsourcing can produce annotations faster and at a fraction of the cost it takes to collect them from experts.

Notwithstanding these differences, most annotation projects assume that a single preferred interpretation or an objective truth exists for each item. But research has shown this assumption to be an idealization at best, both in natural language processing and computer vision. Every large-scale annotation project frequently encounters cases on which humans **disagree**. In some cases, these disagreements are due to misunderstandings or poorly specified annotation schemes. However, in many cases, the interpretation is inherently ambiguous or unclear (Basile et al., 2021). For example, for anaphoric or coreference annotation, Poesio et al. (2007) have discussed **justified sloppiness** in anaphoric reference, as illustrated in example (1).

(1)
```
3.1    M: can we .. kindly hook up
3.2     : uh
3.3     : engine E2 to the boxcar at ..
          Elmira
4.1    S: ok
5.1    M: +and+ send it to Corning
5.2     : as soon as possible please
6.1    S: okay
        [2sec]
```

```
7.1     M: do let me know when it gets
           there
8.1     S: okay it'll /
8.2      : it should get there at 2 AM
9.1     M: great
9.2      : uh can you give the
9.3      : manager at Corning instructions
           that
9.4      : as soon as it arrives
9.5      : it should be filled with
           oranges
10.1    S: okay
10.2     : then we can get that filled
```

In this exchange, it is not clear whether the pronoun *it* in 5.1 (in red) refers to *the engine E2* that has been hooked up to *the boxcar at Elmira* or to the boxcar itself—or indeed whether the distinction matters at all. It is only at utterance 9.5 that we get evidence that *it* probably refers to *the boxcar at Elmira*, since only boxcars can be filled with oranges. Evidence that subjects disagree in such cases has been discussed in several studies (e.g., Poesio and Artstein, 2005; Poesio et al., 2006), and similar cases of disagreements due to justified sloppiness exist in all large-scale anaphoric annotation projects (Versley, 2008; Recasens et al., 2011; Yang et al., 2011; Pradhan et al., 2012).

Indeed, disagreements are frequent in all areas of NLP and in all large-scale annotation projects. The NLP community has realized from the start that it makes no sense to consider gold targets as objective truth in applications such as machine translation, summarization, and natural language generation, where human creativity plays a role and has developed specialized training and evaluation methods for such applications. Recently, the field has tackled classification tasks that involve labelling text according to inherently subjective judgments, such as sentiment analysis (Kenyon-Dean et al., 2018) or offensive language detection (Basile, 2020). It would be clearly misguided to rely on gold labels for training or evaluation in such tasks, as doing so would set one subjective interpretation over all alternatives. Disagreements in interpretation have also been found in annotation projects, such as natural language inference, that ask annotators to make complex judgements (Pavlick and Kwiatkowski, 2019). But disagreements in interpretation are not limited to these complex cases; in fact, they are commonly found even in annotation projects concerned with what might have been thought of as objective and "simple" aspects of language, from part-of-speech tagging (Plank et al., 2014b) to wordsenses (Passonneau et al., 2012) and semantic role labelling (Dumitrache et al., 2019).

The extent of this disagreement varies depending on the complexity and genre of the task, but it can be substantial. For anaphoric annotation, an analysis of the *Phrase Detectives* corpus showed that 64.3% of its data instances contain disagreements.[1] Of those, 12.6% are due to ambiguity, with ambiguous instances making up about 9.0% of the data (Poesio et al., 2019). Recasens et al. (2011) found disagreement due to ambiguity in 12% of the markables in the ANCORA corpus, while Poesio and Artstein (2005) found disagreements in 42% of the markables in the ARRAU dialogue corpus when annotating full anaphoric references. Pradhan et al. (2012) report that for 31% of the disagreements in the standard

1. The *Phrase Detectives* corpus can be found at http://www.phrasedetectives.org/

corpus for coreference resolution, ONTONOTES, the disagreement was caused by linguistic ambiguity. Other aspects of semantic interpretation like lexical disambiguation or semantic role assignment appear to be at least as complex (Passonneau et al., 2012).

(a) (b) (c) (d)

Figure 1: Examples from the LabelMe dataset (Russell et al., 2008)

The assumption that an objective true class exists for all items and that the gold label for each item represents its objective true class has proven an idealization in CV as well. Different coders have assigned different, equally plausible labels to the same items in many widely used crowdsourced computer vision datasets. Consider, for instance, the task of object identification in images. Examples (a), (b), and (c) in Figure 1, discussed in Rodrigues et al. (2017), are from the LabelMe dataset (Russell et al., 2008). Due to the overlap between labels, coders' judgements are highly subjective. The gold label for (a) is "inside city," and one annotator chose that label, but two other annotators chose "tall building." The gold label for (b) is "street." Again, this was produced by one annotator, but two others chose "inside city." The same is true for (c). For (d), none of the annotators chose the gold "street." Instead, all chose "inside city." Clearly, in all these cases, the annotator labels are acceptable, even if they differ from the gold standard. The problem of disagreement among coders, including experts, on the classification of noisy image data arises in CV applications ranging from astronomical classification (Smyth et al., 1994) to the classification of medical images (Raykar et al., 2010) and various others (Sharmanska et al., 2016; Rodrigues and Pereira, 2018; Firman et al., 2018).

Possibly the most widely adopted approach to dealing with disagreements in crowdsourced data is a **source-filter model**, i.e., to assume that there exists a single objective truth that is merely obfuscated by the disagreements, and to use an **aggregation method** over the noisy annotations to find the true "gold" label, a latent parameter (Dawid and Skene, 1979; Carpenter, 2008; Whitehill et al., 2009; Hovy et al., 2013; Passonneau and Carpenter, 2014; Paun et al., 2018). But the presence of systematic disagreements raises serious questions about the basic assumption underlying this approach, i.e., the existence of a single, objectively true label. A number of alternative approaches have therefor been proposed (Sheng et al., 2008; Beigman-Klebanov and Beigman, 2009; Cohn and Specia, 2013; Plank et al., 2014a; Aroyo and Welty, 2015; Guan et al., 2017; Jamison and Gurevych, 2015; Sharmanska et al., 2016; Rodrigues and Pereira, 2018; Li et al., 2019; Uma et al., 2020; Fornaciari et al., 2021). At one extreme, researchers have proposed excluding items on which there is substantial disagreement as bad examples, at least from the test set (e.g., (Beigman-Klebanov and Beigman, 2009)), but possibly also from training. At the other extreme, researchers have argued that "disagreement is signal, not noise" (Aroyo and Welty,

2015)—i.e., that disagreements provide useful information for learning. Various models have been proposed to leverage all of the information provided by annotators, including information about disagreements (Plank et al., 2014a; Aroyo and Welty, 2015; Jamison and Gurevych, 2015; Sharmanska et al., 2016; Uma et al., 2020; Fornaciari et al., 2021). Some of these models do not rely on gold labels at all (Sheng et al., 2008; Sharmanska et al., 2016; Guan et al., 2017; Rodrigues and Pereira, 2018; Firman et al., 2018; Peterson et al., 2019; Uma et al., 2020).

This last view, which advocates for disagreement as a signal, is reminiscent of proposals to improve the generalization abilities of machine learning models through a process known as **distillation**. In distillation, a "student network" is trained using as target the soft output distribution of a "teacher network" (Hinton et al., 2015; Furlanello et al., 2018; Clark et al., 2019). Some distillation work has demonstrated that student networks thus trained (Born Again Networks) can outperform the original teacher networks. Going further, Peterson et al. (2019) have demonstrated that using the distribution extracted from annotations as a soft target results in even better generalization ability than distillation for some image classification tasks.

In this paper, we comprehensively survey the evidence for disagreements on the judgments required to train AI systems. We also survey the range of approaches that have emerged in computational linguistics and computer vision that aim to surpass the simplistic way of dealing with disagreement adopted in most AI approaches to supervised learning. We do not simply review these methods for learning from disagreement; we also use some of the key datasets providing evidence of disagreement to compare them to each other. Furthermore, we compare them to learning from gold labels, assessing the conditions under which each method is effective. In the process, we address the following questions:

- ***RQ1****: What is the most appropriate way of evaluating a model using datasets that provide multiple annotations for each item if we do not assume that every item can be interpreted only one way?*
- ***RQ2*** *is best broken in two sub-questions:*
  (a) *What is the evidence that using information from the crowd annotations helps to build better models than learning from gold labels only?*
  (b) *In case the answer to* ***RQ2a*** *is positive, what is the best way to leverage crowd information in addition to gold labels?*
- ***RQ3*** *is best broken in two sub-questions:*
  (a) *Can methods for learning from disagreement that do not assume the existence of a gold truth (and that do not rely on manual adjudication) perform similarly or better than methods that rely on gold labels?*
  (b) *Which of these methods achieve the best results?*
- ***RQ4*** *is best broken into two parts:*
  (a) *To what extent do the answers to* ***RQ2*** *and* ***RQ3*** *depend on the answer to* ***RQ1****, i.e., on the evaluation method?*

(b) *To what extent do the answers **RQ2** and **RQ3** depend on the task and the characteristics of the annotators of the dataset? Which characteristics matter?*

RQ1 alludes to the fact that if we question the existence of a single gold label for each item, then evaluating models based on how well they replicate that gold label does not make much sense. While proposals for alternative, "soft" evaluation metrics (i.e., not dependent on that assumption) exist (Firman et al., 2018; Dumitrache, 2019; Peterson et al., 2019), in practice, most studies still use "hard" evaluation metrics such as accuracy. We are not aware of any previous attempt to systematically evaluate models with both "hard" and "soft" evaluation metrics. Addressing this gap is one of the main contributions of this study.

Some papers have surveyed the impact of disagreements on learning from crowdsourced labels, but none as comprehensively as the present one. The paper that comes closest is by Jamison and Gurevych (2015), who examined the impact of infusing classifiers with information from annotators, but only considered two of the approaches reviewed here, hard filtering and the soft-labelling approach proposed by Sheng et al. (2008), which we will refer to as **srel** (for Sheng REpeated Labelling). They compared these two methods to training with gold labels and labels aggregated via majority voting, finding that soft labelling did not outperform training with gold or aggregated labels for their datasets using their classifiers and hard metrics. Our study examines a great many methods beyond those explored by Jamison and Gurevych. Additionally, we investigate the effect of disagreement on the success of each training method, evaluating model performance using soft metrics in addition to hard metrics. Furthermore, we use current, state-of-the-art baselines based on neural networks as our starting point. As Jamison and Gurevych's paper remains one of the few contributions to this area, we used two of their datasets to compare our findings with theirs.

This survey paper is structured as follows. Section 2 reviews the evidence that humans disagree on many tasks in language and image interpretation and presents the datasets we employed (these are available as supplementary materials). Section 3 reviews the most important approaches to learning from disagreement. Section 4 discusses possible answers to **RQ1**, i.e., how models trained using disagreement methods can be tested on datasets containing multiple and possibly disagreeing judgments. Section 5 describes the experimental design used to answer our research questions. It discusses the models used for each dataset and the methods used to train them. Section 6 contains the results of our experiments. Section 7 breaks down the results task by task, drawing attention to the effect of several aspects of each task on the results. Section 8 contains a general discussion of the results and their insights into our research questions. Finally, Section 9 summarizes the conclusions of this study.

2. Disagreements in NLP and CV: Evidence and Resources

As mentioned above, an extensive body of literature provides evidence of the extent to which humans disagree on many aspects of interpretation in NLP and CV. In this section, we discuss some of this evidence in greater detail, focusing on studies that also created datasets that preserve these disagreements. These datasets form the basis of our experimental evaluation in Sections 5–7.

We chose tasks with the aim of using deep learning methods that reflect the current state of the art to improve upon previous experimental studies on learning from disagreement, in particular Jamison and Gurevych (2015). This constraint restricted us to tasks for which datasets large enough to train such models exist; our rule of thumb was to consider only datasets of at least 1,000 items. The one exception was in the task of recognizing textual entailment (RTE)/natural language inference, where some key work on disagreements in interpretation has been carried out, in particular by Pavlick and Kwiatkowski (2019). We decided to use for this task the dataset by Snow et al. (2008), which only consists of 800 items but has been ubiquitous in research on crowdsourcing and aggregation and was used by Jamison and Gurevych (2015).

The NLP tasks we selected include part-of-speech (POS) tagging, which led to the creation of the Gimpel corpus (Plank et al., 2014b,a; Jamison and Gurevych, 2015)); information status (IS) classification, a simplified version of the anaphoric interpretation task studied in some of the early work on disagreements (Poesio and Artstein, 2005; Poesio et al., 2006) and for which we were able to leverage the largest NLP corpus providing multiply annotated data, *Phrase Detectives* (Poesio et al., 2019); (medical) relation extraction (MRE), extensively studied in the CrowdTruth project (Aroyo and Welty, 2015; Dumitrache, 2019; Dumitrache et al., 2019), which resulted in the creation of several datasets including the one used in this study (Dumitrache et al., 2018b); and recognizing textual entailment (RTE), which led to the development of the Snow et al. corpus used, for example, in Snow et al. (2008); Jamison and Gurevych (2015). For CV, we focused on image classification (IC), using two important datasets created to study learning from disagreement: the LabelMe corpus (IC-LABELME), a crowdsourced version of which was created by Rodrigues and Pereira (2018), and the CIFAR-10H corpus, recently crowdsourced by Peterson et al. (2019). This section briefly discusses each of these tasks, their respective datasets, and the evidence of disagreement they provide. We used a standardized format for the task descriptions to facilitate comparison and summarize the characteristics of the datasets in Section 2.6.

2.1 Part-of-Speech tagging

POS tagging is the task of assigning part-of-speech tags such as noun or verb to every word in a text. It is thought to reflect a very basic aspect of human lexical or syntactic competence, and we would therefore expect little or no disagreement in the judgments of coders asked to carry out this type of annotation. But in fact, one of the best-known studies in the area of learning from disagreements, Plank et al. (2014b), was motivated by the observation that annotators systematically disagree even on such supposedly simple linguistic tasks as this one. Plank et al. found systematic disagreements between, for example, adpositions (ADP) and particles (PRT), as in *get* <u>*out*</u>; adjectives (ADJ) and nouns, as in <u>*stone*</u> *lion*; and adjectives and adverbs (ADV), e.g., in *see you* <u>*later*</u>. They found the same disagreements among experts and non-experts, and across text types. Plank et al. investigated the nature of these disagreements, finding that while some disagreements were a result of annotation error, others were evidence that the category of certain items was linguistically debatable. They further discovered that making the annotation guidelines increasingly more detailed did not eliminate these latter errors or "hard cases" (Plank et al., 2014b). They thus

hypothesized that these disagreements were a result of label uncertainty and could be used to inform the learning process.

The dataset The analysis by Plank et al. (2014b) was carried out as part of the creation of one of the best-known datasets for research on learning from disagreement, and the first dataset chosen for this study. This dataset—henceforth, mostly abbreviated as POS—builds upon the Gimpel et al. (2011) corpus of POS labels for Twitter posts with the crowdsourced labels provided by Hovy et al. (2014). Plank et al. mapped the Gimpel tags to the universal 12-tag set (Petrov et al., 2011), using these tags as gold labels, and collected at least 5 crowdsourced labels per token from 177 annotators. The dataset consists of over 14k examples and was previously used in Plank et al. (2014a) and Jamison and Gurevych (2015). We used the data released by Plank et al. (2014a) as a development set.[2]

Annotations and annotators The size of the crowd employed to collect judgments is important(Snow et al., 2008). A number of studies (Poesio and Artstein, 2005; Dumitrache, 2019; Peterson et al., 2019) have shown that the number of annotations collected is also of key importance for studying disagreement. For instance, Poesio and Artstein (2005) have shown that what they call **implicit ambiguity**—the ambiguity emerging from disagreements among annotators, rather than from annotators explicitly marking items as ambiguous—only starts to emerge for the task of anaphoric annotation when at least 5 annotations per item are collected. (The precise number of annotations appears to depend on the task.) Each item in the Gimpel dataset was annotated 5 times, apart from 946 items with a much greater number of annotations—these were most likely tutorial items (Gimpel et al., 2011). The percentage of items annotated by each coder ranges from 2.64% to 5.29%. Given that there are 12 possible categories, the ratio number of coders to possible categories (the coder:label ratio) is 5:12 or 0.416.

Also important is the level of agreement between these annotators; the average item-observed agreement, computed using the Fleiss multi-annotator version of the kappa statistic (Fleiss et al., 2004) (henceforth: κ), is 0.725 overall and 0.706 excluding tutorial items. We also note the performance of the annotator with respect to the gold label as a way to measure the degree of alignment between the experts and the annotators. This measurement is an indicator of how much the gold stands apart from the crowd. We use accuracy for this measure and in this dataset, the average accuracy per annotator in the POS dataset is 67.81%, with over 38.98% of coders falling below this average. Only about 29% of annotators have a near-gold performance, achieving 75% or more accuracy with respect to gold labels. We highlight this point because noise is a factor affecting the ability of models to learn from crowds, as discussed in Sections 7 and 8.

Quality of aggregated labels We also measure the accuracy of aggregated labels with respect to the gold as it indicates how much the crowd consensus aligns with the expert label. The accuracy of aggregated labels with respect to gold labels indicates how well the crowd consensus aligns with the expert label. There is substantial disagreement in this dataset: 48.09% of the items received annotations assigning them to more than one category. Majority voting accuracy with respect to gold labels is 79.69%; the Dawid and Skene (1979) and MACE aggregation methods (Hovy et al., 2013) discussed later produce

2. Plank et al.'s data can be found at `http://lowlands.ku.dk/results/`

labels that are 79.13% and 79.83% accurate, respectively, when accuracy is determined by gold labels.

2.2 (Anaphora and) Information Status Classification

Possibly the first type of disagreement systematically studied in NLP was disagreement on anaphoric annotation (coreference), already identified by the previously mentioned studies by Artstein and Poesio (Poesio and Artstein, 2005; Poesio et al., 2006), further evidence for which was unearthed as part of the annotation of virtually every modern corpus of anaphoric information: ANCORA (Recasens et al., 2011), ARRAU (Poesio and Artstein, 2008; Uryupina et al., 2020), ONTONOTES (Pradhan et al., 2012), *Phrase Detectives* (Poesio et al., 2019), the Potsdam Commentary Corpus (Krasavina and Chiarcos, 2007), the Prague Dependency Treebank (Nedoluzhko et al., 2016), and TUBA/DZ (Versley, 2008). Anaphora is a more complex task than POS tagging, but it is still considered a basic aspect of language interpretation. Yet the previously mentioned researchers found disagreements on the anaphoric interpretation of between 12% and 40% of all mentions depending on the genre and the range of anaphoric phenomena considered (Poesio et al., 2019). Besides the examples of ambiguity as to the antecedent of an anaphoric expression discussed in the Introduction, this research found subjects disagreeing as to whether nominal form *it* was anaphoric or expletive (as in *when she [Alice] thought it over afterwards, it occurred to her that she ought to have wondered about this ...*); whether a nominal introduced a new entity or referred to an old one; and more complex cases of ambiguity related to the antecedent, e.g., in cases of reference to "split antecedent" plurals and discourse deixis (Poesio et al., 2019; Recasens et al., 2011).

Because of the complexity of adapting models of learning from disagreement to full anaphora/coreference resolution, in this study we only looked at disagreements on a simplified form of the task, information status classification (IS), which involves identifying the information status of a noun phrase, i.e., whether that noun phrase refers to a new entity or to an entity that has already been introduced.

Dataset There are different annotation schemes for annotating information status (Prince, 1981, 1992; Nissim et al., 2004; Riester et al., 2010). The dataset we used, and that we call PDIS here, is extracted from the *Phrase Detectives* 2 corpus for coreference resolution (Poesio et al., 2019),[3] which used a simplified, binary definition of the IS derived from the annotation scheme used in *Phrase Detectives*. In PDIS, only markables classified as introducing a new entity (discourse new, DN) or as referring to a previously introduced entity (discourse old, DO) are considered. Markables classified as expletives or as predicative are not considered, and information about coreference chains is ignored.

To our knowledge, the *Phrase Detectives* 2 corpus is the largest NLP corpus with multiple annotations. It consists of a total of 542 documents containing 408k tokens and about 108k markables. Of these, 497 documents were used for training and development and 45 were used for testing. These documents were annotated by over 1,828 annotators producing at least 8 annotations per markable. There are no expert annotations for the 497 training/development documents, but the 45 documents for testing contain both expert and

3. The *Phrase Detectives* 2 corpus is freely available from the LDC and from `https://github.com/dali-ambiguity`.

crowd annotations. The training, development, and test data respectively contain 97,040, 4,753, and 5,855 markables.

Annotations and annotators The full *Phrase Detectives* 2 corpus contains a total of 2,235,664 judgments, for an average of 20.6 annotations/validations per item. After restricting the judgments to only the binary DN/DO labels and excluding validations (Poesio et al., 2019), we were left with an average of 11.87 annotations per item for the PDIS binary subset (or 7.01 if only one annotation was counted for each annotator). The average observed agreement per item is 0.809. Each coder annotated 413.75 items on average, and the average coder accuracy is 78.13%. At least 71% of coders are at least 75% accurate.

Quality of aggregated labels In PDIS, considering only the subset of data for which gold labels are available, the labels aggregated using majority voting are 89.54% accurate, whereas the labels aggregated using Dawid and Skene (1979) and MACE are 98.14 % and 97.89 % accurate, respectively.

2.3 Relation Extraction and Frame Disambiguation

Another aspect of semantic interpretation for which there is extensive evidence of disagreements among annotators is relation extraction—the task of deciding, given two mentions and a segment of text (clause or sentence), whether that segment expresses one among a fixed number of relations between the entities referred to by those mentions. Relation extraction was one of the two most extensively studied tasks in the CrowdTruth project (Aroyo and Welty, 2015; Dumitrache et al., 2018a, 2019). Aroyo and Welty (2015) discuss examples encountered in projects for crowdsourcing medical relation extraction such as (2):

(2) GADOLINIUM AGENTS used for patients with severe renal failure show signs of NEPHROGENIC SYSTEMIC FIBROSIS.

Annotators asked to label the relation between the underlined pairs with one of the unified medical language system UMLS relations systematically disagreed on whether pairs such as the one in the example were instances of the **cause** (strict sufficient causality) relation or the **side-effect** (possibility of a condition arising) relation. Again, neither experts nor novice annotators were able to systematically make the distinction.

Two types of relation extraction were studied in CrowdTruth: medical relation extraction (MRE), the application to medical texts, and frame disambiguation, the version of the task in which the repertoire of relations is provided by FrameNet (Dumitrache et al., 2019). We focus on MRE in this paper.

Dataset We used the medical relation extraction (MRE) dataset from Dumitrache et al. (2018b). They created a dataset of 3,984 English sentences extracted from PubMed article abstracts for medical relation extraction centered on two main relations, the **cause** and **treat** relations, that have been processed with disagreement analysis to capture ambiguity. The sentences were sampled from the set collected by Wang and Fan (2014) using distant supervision (Mintz et al., 2009).

Dumitrache et al. (2018b) collected expert annotations for a randomly sampled set of 975 sentences from the distant supervision dataset, with each sentence being annotated by a single expert. The annotation task involved deciding whether or not the UMLS seed relation

discovered by distant supervision did in fact hold between two highlighted terms in a given sentence (Dumitrache et al., 2018b). The crowdsourcing was carried out using so-called **disagreement-aware crowdsourcing** (Aroyo and Welty, 2015): For every sentence, the crowd was asked to choose any number of relations from 14 possible relations, including **other** and **none**, applicable to the highlighted terms in the sentence.[4]

For our experiments, we only considered the **cause** relation. Like Dumitrache et al. (2018b), we framed the task as a binary classification. The gold label for each sentence given the highlighted terms was 1 if the expert agreed that the two entities stood in a **cause** relation, 0 otherwise. Similarly, for each annotator who annotated the sentence, the assigned label was 1 if the annotator selected **cause** amongst his/her choices, 0 otherwise.

Annotations and annotators Each of the 975 sentences was annotated by at least 15.3 annotators (a minimum of 15 and a maximum of 30). On average, each coder annotated 5% of the items (a minimum of 0.1% and a maximum of 43.58%) and the average annotator accuracy was 76.1% (minimum of 0% and maximum of 100%); 58% of the annotators were at least 75% accurate. The observed agreement per item was 0.857.

Quality of aggregated labels The majority voting label was aggregated by counting the number of workers who selected the *cause* relation as a valid relation for the sentence. Labels aggregated using majority voting were 74.6% accurate with respect to the gold labels. Labels aggregated using Dawid and Skene (1979) and MACE were 76% and 76.61% accurate, respectively. Dumitrache et al. (2018b) also provide labels aggregated using the CrowdTruth approach (discussed in section 3). These labels are 80.51% accurate with respect to the gold labels.

2.4 Recognizing Textual Entailment/Natural Language Inference

Another aspect of language interpretation for which there is systematic evidence of disagreement among subjects is recognizing textual entailment (henceforth, RTE) (Dagan et al., 2005).[5] Recognizing textual entailment/natural language inference is deciding whether the proposition conveyed by a text (the hypothesis h) can be inferred from another proposition (the premise p) (Dagan et al., 2005). In NLP this task is typically formulated as a binary classification task in which a pair p/h is classified as true if the hypothesis can be inferred from the premise, false otherwise.

RTE attempts to model what is arguably the foundation of semantics (Cooper et al., 1996), but it has proven hard for humans to agree on RTE judgments. Lalor et al. (2018) discuss examples like (3), in which it is not clear whether the hypothesis that the child plays/intends to play with the balloon follows from the premise that he's reaching for it and laughing.

(3) a. *Premise:* A young boy in a beige jacket laughs as he reaches for a teal balloon.
b. *Hypothesis:* The boy plays with the balloon.

4. The dataset by Dumitrache et al. (2018b) is available from `https://github.com/CrowdTruth/Medical-Relation-Extraction`.
5. The term "natural language inference" is now also used (Bowman et al., 2015; Pavlick and Kwiatkowski, 2019), but we will mainly use the term RTE given that this is the name of the dataset we used.

In a recent and systematic analysis of inherent disagreement on RTE judgments, Pavlick and Kwiatkowski (2019) found that workers disagreed on at least 20% of the p/h pairs they were asked to classify. They also found that a mixture of Gaussian models generalized better to unseen examples than single-component Gaussians, providing additional evidence that unimodal label distributions are at best an idealization—human disagreement is a reproducible signal. Pavlick and Kwiatkowski (2019) have further analyzed the uncertainty from the model output distribution, showing that it differs from the human distribution; this work has inspired a recent study on disagreement for RTE/NLI (Nie et al., 2020). Nie et al. have shown that NLI models achieve near perfect accuracy on high-agreement instances but fall to random levels when there is low human agreement.

The dataset To study the effect of disagreements on learning RTE models, we used the classic PASCAL RTE-1 challenge dataset (Dagan et al., 2005), which contains 800 text-hypothesis pairs (Dagan et al., 2005). Crowdsourced annotations for this corpus were later collected by Snow et al. (2008); 164 annotators produced 10 annotations for each sentence pair. Each sentence pair also received a gold label. We chose this dataset because it allowed us to compare our results with those of other researchers who have studied disagreement in RTE, in particular Jamison and Gurevych (2015).[6] It is also substantially larger than the datasets produced by Lalor et al. (2018) and Pavlick and Kwiatkowski (2019).

Annotations and annotators In PASCAL RTE-1, each item received exactly 10 annotations from one of the 164 coders. This is a binary classification dataset, and the coder:label ratio is 10:2 (5). The average observed agreement across all items, computed using κ (Fleiss et al., 2004), is 0.629.

Each coder annotated between 2.5% and 100% of the items, 6.09% on average. The average accuracy per annotator was 83.70%, and only 35.37% of annotators fell below this average—82.93% were at least 75% accurate.

Quality of aggregated labels There was a substantial amount of pre-aggregation disagreement in this corpus, much higher than with the POS dataset: 91.88% of the items had more than one interpretation. However, the alignment between aggregated (silver) labels and gold labels was much higher than with POS. Majority voting aligned with the gold label in 90.25% of cases, while using the Dawid and Skene (1979) and MACE aggregation methods produced labels that aligned with the gold labels in 92.88% and 92.63% of cases, respectively.

2.5 Image Classification

Image classification is a very general term for the task of assigning an image to the category that best describes it among a fixed set; the available categories depend on the application. Historically, it has given rise to more research on learning from disagreement than perhaps any other area of AI. Image classification has provided the motivation for developing methods for aggregating multiple expert-produced labels, particularly for medical images (Dawid and Skene, 1979; Smyth et al., 1994; Whitehill et al., 2009). More recently, researchers working on applications of this type have started to develop methods that learn

6. The dataset from Snow et al. (2008) is available from `http://sites.google.com/site/nlpannotations`

classifiers directly from the labels produced by the crowd (Raykar et al., 2010; Albarqouni et al., 2016; Guan et al., 2017; Rodrigues and Pereira, 2018; Peterson et al., 2019). We therefore considered it essential to include datasets used in (and often originating from) this type of research in our assessment of methods for learning from disagreement. Specifically, we employed two datasets, both extensively used in the CV literature.

2.5.1 LabelMe

Dataset LabelMe (Russell et al., 2008) is a widely used, community-created image classification dataset where images are assigned to one of 8 categories: **highway**, **inside city**, **tall building**, **street**, **forest**, **coast**, **mountain**, **open country**.[7] Rodrigues and Pereira (2018) collected crowd labels for 10,000 of these images using Amazon Mechanical Turk to engage 59 annotators producing at least one label for each image. In this study, we used this version of LabelMe, henceforth IC-LABELME.

Annotations and annotators Each item in the IC-LABELME dataset was annotated at least once and a maximum of 3 times; the average number of annotations per item is 2.55. With 8 classes for this dataset, the average ratio number of coders to possible categories is 2.55:8, or 0.318. The average item observed agreement, computed using κ (Fleiss et al., 2004), is 0.732.

Each coder annotated from 0.3% to 18.2% of items, for an average of 6.09%. The average accuracy per annotator was 69.16%. While over 38.98% of coders fell below this average, 42.37% were at least 75% or accurate.

Quality of aggregated labels For this dataset, majority voting aggregation produced labels with 76.9% accuracy with respect to the gold labels while Dawid and Skene (1979) and MACE aggregation generated labels with 79.9% and 78.3% accuracy, respectively.

2.5.2 CIFAR10H

Dataset The CIFAR-10 dataset is another state-of-the-art image classification dataset (Springenberg et al., 2015; Graham, 2014; Ghosh et al., 2017; Gastaldi, 2017).[8] In full, it consists of 60k 32x32 color images in 10 categories (**airplane**, **automobile**, **bird**, **cat**, **deer**, **dog**, **frog**, **horse**, **ship**, and **truck**) with 6k images per class. There are 50k training images and 10k test images.

Recently, this dataset has also been used in research into learning from crowdsourced data. In particular, Peterson et al. (2019) collected human annotations for the test portion of CIFAR-10 using Amazon Mechanical Turk, creating the "CIFAR10H" dataset.[9] This dataset consists of 511,400 human categorization decisions over 10k images with an average of 50 annotations per image.

In this paper, we used the CIFAR10H dataset, referred here as IC-CIFAR10H, for training and testing using a 70:30 random split while ensuring that the number of images per class remained balanced as in the original dataset. We also used a subset of the CIFAR-10 training dataset (3k images) as our development set.

7. The LabelMe dataset can be found at `http://labelme.csail.mit.edu/`
8. The CIFAR-10 dataset is available at `https://www.cs.toronto.edu/~kriz/cifar.html`
9. The dataset from Peterson et al. (2019) is available from `https://github.com/jcpeterson/cifar-10h`

Annotations and annotators Each item was annotated an average of 51.1 times (a minimum of 47 and a maximum of 63). Given that there are 10 possible classes, the average coder:label ratio is 5.11. The average observed agreement per item is 0.924.

Peterson et al. (2019) removed annotators with less than 75% accuracy from the dataset, so 100% of the remaining coders were at least 75% accurate, resulting in an average annotator accuracy of 95%.

Quality of aggregated labels Majority voting produced labels with 99.21% accuracy with respect to the gold labels, while aggregating using Dawid and Skene (1979) and MACE aggregation generated labels with 99.27% and 99.24% accuracy respectively.

2.6 A Summary of the Datasets used in this Study

Tables 1 and 2 summarize the key statistics related to the datasets discussed in this section.

Table 1: Annotations and Annotators

	POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
Number of items	14,000	96,305	975	800	10,000	10,000
Number of crowd workers	177	1,741	304	164	59	2,457
Number of labels	12	2	2	2	8	10
Average annotations per item	16.37	11.87	15.30	10.00	2.50	51.10
Median annotations per item	5	10	15	10	3	51
Average number of items annotated per coder	1335.48	381.757	49.08	48.78	431.69	200
Median number of annotations per coder	1276	20	14	20	270	200
Average coder accuracy	0.93	0.82	0.76	0.84	0.69	0.95
Coder accuracy variance	0.003	0.062	0.053	0.015	0.033	0.001
Percentage of coders with accuracy above 0.75	1.00	0.77	0.58	0.83	0.42	1.00
Average observed agreement per item	0.73	0.81	0.86	0.63	0.73	0.92
Average item entropy using raw distribution	0.13	0.38	0.31	0.72	0.10	0.07
Average item entropy (best-performing distribution, BDE)	0.39	0.09	0.31	0.22	0.76	0.07

Table 2: Quality of Aggregated Labels

	POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
Percentage accuracy of MV aggregated labels	0.80	0.89	0.75	0.90	0.77	0.99
Percentage accuracy of D&S aggregated labels	0.79	0.98	0.76	0.93	0.80	0.99
Percentage accuracy of MACE aggregated labels	0.79	0.98	0.77	0.93	0.78	0.99

As the tables show, the datasets differ along a number of dimensions, from the average number of annotations per item to the average number of annotations per coder to the accuracy of coders to the degree of confusion, measured in terms of observed agreement and entropy. We computed two forms of entropy: entropy according to the "ra" soft label for an item (the probability distribution based on the annotators' raw annotations) and **best distribution entropy** (BDE) (the entropy according to the soft label that performed best for a particular dataset). As we will see in later sections, these differences help us understand the differences in performance among the approaches to learning from disagreement studied in this paper.

2.7 Other Tasks

Although we focused on the six tasks (datasets) discussed above, judgment disagreements have been observed by virtually every major annotation project for virtually all language and vision interpretation tasks. In this section we briefly review some of the literature on the presence of systematic disagreement in other aspects of language interpretation.

Syntactic interpretation Alonso et al. (2015) observed that the disagreements noted by Plank et al. (2014b) were largely characteristic of dependency parsing and applied the method proposed by Plank et al. (2014a), which also yielded promising results.

Wordsense disambiguation and supersenses Projects on wordsense annotation gave rise to an early line of research on disagreements in NLP that arose alongside research on anaphoric disagreement. Possibly the best known in this area is the seminal work of Passonneau and colleagues on wordsense disambiguation in the American National Corpus (see, e..g, Passonneau et al. (2012) and Passonneau and Carpenter (2014)). Passonneau et al. (2012) carried out a systematic analysis of disagreements related to different types of words (nouns, verbs, and adjectives), investigating the extent to which disagreements depended on annotator quality, instructions, and context. Further investigations of the practice of wordsense annotation were carried out by Jurgens (2013). Alonso et al. (2016) showed that disagreement arises also in supersense tagging, and they performed experiments using the method developed by Plank et al. (2014b) on English and Danish supersense datasets.

Named entity recognition The other NLP task systematically explored in the CrowdTruth project was named entity recognition (NER) (Inel and Aroyo, 2017). This task was the second application of the methods developed in that project, discussed in Section 5.2. Named entity recognition was also one of the test applications in Rodrigues and Pereira (2018).

Discourse structure More disagreement is to be expected when considering tasks requiring more complex judgments, such as analyzing discourse structure. This intuition was confirmed in early work by Stede (2008). More recently, further evidence has been provided in work on argument structure annotation. The AURC-8 corpus (Trautmann et al., 2019) contains gold-standard annotations for argument component spans derived from crowdsourced labels. As well as disagreement over whether a span is argumentative or not, the starts and ends of argument components are often ambiguous, leading to significant disagreements between annotators. Simpson and Gurevych (2019) used as one of the datasets for testing their sequence-learning from crowds method a subset of the crowdsourced annotations from AURC-8 containing 8000 sentences, each with five judgements from 105 annotators.

Sentiment analysis and other subjective tasks Even more disagreement is to be expected with subjective tasks such as sentiment analysis or hate speech (Basile, 2020). This intuition is confirmed by evidence such as the study by Kenyon-Dean et al. (2018) in support of the annotation of the McGill Twitter Sentiment Analysis corpus. Kenyon-Dean and colleagues found that over 30% of the instances in the corpus were "controversial" or "complicated" cases about which annotators disagreed. Akhtar et al. (2019) experimented

with partitioning the annotators in hate speech datasets into clusters reflecting more uniform subjective judgments in order to achieve increased inter-annotator agreement.

2.8 Sources of Disagreement

While all disagreement results in label uncertainty, some of the examples of disagreement discussed in the Introduction are the results of ambiguity and/or subjectivity and are hence intrinsic to a given task (for instance, the example of anaphoric ambiguity). Others are a result of annotator or annotation interface errors or problems with the annotation scheme and introduce noise to the data. The six datasets employed in this paper are characterized by different forms of disagreement; understanding the nature and sources of the disagreements found in a dataset would thus appear to be an essential prerequisite if we are to properly harness disagreement in building machine learning models for that dataset. Several annotation projects have attempted to classify possible sources of disagreement—with varying degrees of success (see, e..g, Poesio et al. (2019)). In this section we discuss the sources of disagreement that occur in our six datasets, attempting to classify them despite the difficulty of making some of the distinctions.

2.8.1 Errors and Interface Problems

Traditionally, disagreement has been viewed as the result of **annotator errors**: mistakes or slips made by the annotator. Several annotation projects have highlighted this source of disagreement. Indeed, Nedoluzhko et al. (2016) found that 15% of annotator disagreement was as a result of annotator errors, Pradhan et al. (2012) attributed 25% of the disagreements in ONTONOTES to annotator error, and Plank et al. (2014b) found that while the ratio of noise:genuine ambiguity differed depending on the level of confusion of label pairs, annotation errors were responsibly for 30% of the disagreement on difficult items. Disagreement due to annotator error, while not informative about the task itself, provides information about the reliability of the annotators, their level of attention, or their level of understanding of the task.

Errors resulting in disagreement could also be the result of **interface problems** or limitations. In coreference annotation, for example, errors in the markable extraction process (i.e., incorrectly defined span boundaries for markable noun phrases) often introduce disagreements: Annotators, unable to select the appropriate span, either select a preceding antecedent in the same chain, a span which is a subset of the correct span, or annotate the markable as problematic. These differing judgements lead to unnecessary disagreements. Consider the following sentence from the *Phrase Detectives* corpus, where the (automatically) extracted markable is in bold font and surrounded by square brackets:

> "Once upon a time there was a dear little girl who was loved by everyone who looked at her but [**most of all by her grandmother**]"

most of all by her grandmother is not a valid markable; as a result, annotators disagree on the most suitable label. A majority of the annotators marked the markable as DN, "discourse new," while the other annotators marked it as "predicative." Our analysis of a sample of documents in the corpus showed that interface limitations and problems accounted for a

majority of the disagreement in the validated *Phrase Detectives* corpus (Poesio et al., 2019). In ONTONOTES, another coreference corpus, interface and annotation scheme limitations account for 43% of the disagreements. As with annotator errors, disagreements resulting from interface errors are not informative about the tasks but are useful information about the annotation project.

2.8.2 Annotation Scheme

Incomplete, imprecise, or vague annotation schemes may also result in annotator disagreements. Such annotation schemes may contain ill-defined classes or overlapping classes and/or may not cover all items, leaving an annotator unsure as to the best label for an item (Nedoluzhko et al., 2016). As discussed in the Introduction, the annotation scheme used in IC-LABELME (Russell et al., 2008) is a prime example of a scheme containing overlapping class names and vague descriptions. For instance, the classes **inside city**, **street**, and **tall building** are not mutually exclusive, so an annotator forced to choose only of out of of the three will likely make an arbitrary decision. Figure 2 shows three images with buildings, each assigned a different gold label. Such examples show that even "gold" labels for such items merely reflect the biases of the expert annotators, not an objective truth.

(a) **inside city** (b) **street** (c) **tall building**

Figure 2: Examples showing similar images from LabelMe captioned with their gold labels (Russell et al., 2008; Rodrigues and Pereira, 2018)

It is therefore not surprising that annotators would disagree on the interpretation of such items. Figure 3 shows the confusion matrix between majority voting consensus and gold labels for the crowdsourced LabelMe dataset collected by Rodrigues and Pereira (2018). Images classified as **inside city** by the gold label were assigned to the category **tall building** by the majority 22% of the time, while images classified by the gold label as **street** were assigned the label **inside city** by a majority of annotators 26% of the time.

This is unsurprising considering Figure 2: **Street**s can have **tall building**s and are often located **inside city**. It is also understandable that images classified as **open country** by their gold labels are assigned the class **mountains** by the majority 23% of the time; this is justifiable because **open country** sometimes contains **mountain**s. Figure 4 gives an illustration of this.

Similar overlap exists among other label pairs. Merging the 8 fine-grained categories of LabelMe into 3 categories—(1) **coast**, (2)**inside city + street + tall building**, and (3) **forest + mountain + open country**—results in majority voting aggregated labels that accord with gold labels in 95% of cases, 18% more than the when the labels are left unmerged.

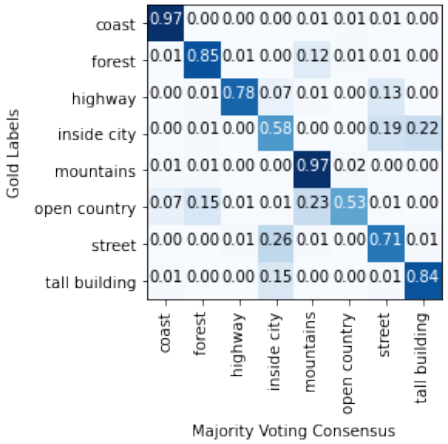
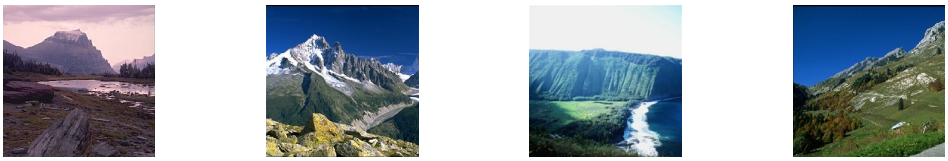


Figure 3: Confusion matrix between gold labels and majority voting consensus for LabelMe



(a) **open country** (b) **mountain** (c) **open country** (d) **mountain**

Figure 4: More examples showing similar images from LabelMe, each assigned a different gold label (Russell et al., 2008; Rodrigues and Pereira, 2018)

2.8.3 Ambiguity

At the beginning of the corpus-driven revolution in NLP it was assumed that all disagreements in annotation were due either to errors, interface problems, or poorly defined annotation schemes. But in fact, many disagreements among annotators are due to **ambiguity**: the fact that a number of expressions can be interpreted in semantically distinct ways in a given context (Poesio, 2020). Ambiguity is not a consequence of a poor annotation scheme but of the inherent complexity of understanding and interpretation. Several studies have found genuine ambiguity to be a leading source of disagreement. The example of anaphoric ambiguity in the Introduction comes from a seminal work studying disagreement as evidence for ambiguity, the paper by Poesio and Artstein (2005) on annotation of anaphora in dialogue data. Poesio and Artstein employed 18 students to annotate the same segments of a dialogue from the TRAINS corpus by selecting *all* valid antecedents for every markable expression the annotator perceived to be ambiguous. They found that at least 10% of the 72 markables annotated were marked **explicitly** ambiguous by at least one annotator (Poesio and Artstein, 2005). They also found cases of **implicit ambiguity**, where markable items were not marked as ambiguous by annotators but different annotators chose different, equally valid labels. Our analysis of some documents from the *Phrase Detectives* corpus showed similar results. We found that while a majority of the disagreements were the re-

sult of interface problems, 9.1% could plausibly belong to more than one coreference chain (Poesio et al., 2019).

Further evidence of disagreement due to ambiguity can be found in the POS dataset used in this paper. Plank et al. (2014b) analyzed the inter-annotator disagreements in this corpus and demonstrated that some disagreements were consistent across domains and languages, and certain label pairs were more confusing than others (Plank et al., 2014b). They employed expert linguists to annotate 880 items from the Gimpel et al. dataset, finding that a majority of the disagreements for certain label pairs stemmed from linguistically debatable cases (Plank et al., 2014b). For example, Plank et al. found that all the NOUN-PRON disagreements were always linguistically debatable cases; the same was true for 70% of the ADP-ADV disagreements. Figure 5 shows the the result of Plank et al.'s analysis of

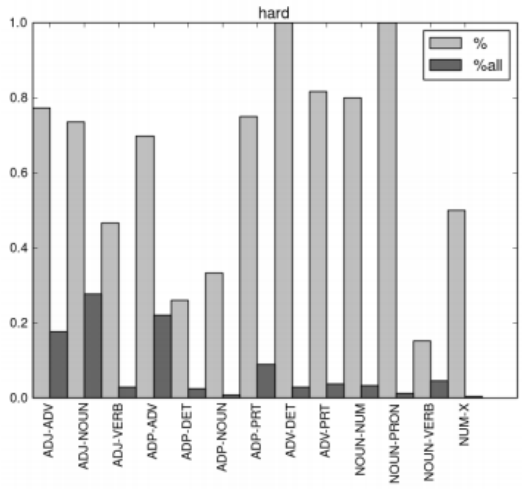


Figure 5: Figure showing the proportion of hard cases that make up 880 POS items (dark gray) and the proportion of these hard cases that are linguistically motivated (light gray) (Plank et al., 2014b)

the disagreement involving these label pairs; the dark gray bars show the relative occurrence of a pair confusion in the dataset, while light gray bars show the proportion of disagreement that is due to linguistic ambiguity. These results were further validated when 10 linguistic faculty members were asked to select the correct label for 10 items in the dataset; in 8 out of 10 cases, these experts disagreed on on the right tag.

It should be noted that the presence of ambiguity in a text is not always **nocuous** (Yang et al., 2011)—i.e., not being able to recover the intended interpretation of some expressions need not be problematic (Poesio and Artstein, 2005; Poesio, 2020). For instance, in example (1), the fact that pronoun *it* is ambiguous in context does not affect the listener in that the two interpretations are equivalent: The ambiguity is innocuous. But Yang et al. (2011) discuss several cases in which the ambiguity of an expression may be problematic.

2.8.4 Item Difficulty

In the cases of ambiguity discussed in 2.8.3, the interpretations of an expression are clear even though it is not clear from the context which interpretation is intended. There are cases, however, where disagreement is caused by the fact that it is not clear what the interpretation of an item is, if any. We grouped these cases in a distinct category that we call **item difficulty** disagreement.[10] While difficulty in this sense is encountered in all annotation projects, item difficulty is a leading cause of disagreement for two of the datasets studied in this paper, the RTE dataset and the CIFAR dataset.

Zaenen et al. (2005) have noted that the PASCAL dataset contains examples that do not fit a clearly defined inference pattern, making them problematic to categorize. As an illustration, consider the randomly selected high-disagreement (**polarizing**) items in RTE shown in Table 3. We can observe from Table 3 that the items on which annotators disagree contain convoluted premises (2 and 9), convoluted hypotheses (3 and 8), or require extra-textual information that annotators need to supply based on their real-world knowledge (1, 4, 5, 6, 7, 10). We classified these cases of disagreement in RTE as being caused by (general) difficulty rather than ambiguity because the entailment relation cannot be both *true* and $false$; it is just that it is not clear what the correct value is. For example, item **7** is difficult because it is unclear whether or not the newly appointed editor-in-chief has assumed his role; however, the hypothesis "*Al Jumhoria is the Iraqi Ambassador to India*" either follows from the premise or doesn't; it cannot do both. These polarizing examples stand in contrast to the non-polarizing (perfect agreement) instances in RTE shown in Table 3.[11]

The nature of difficulty in CIFAR can be illustrated by contrasting the example images in Figure 7, about which the annotators perfectly agree, with the example images in Figure 6, for which observed agreement is less than 0.3. The images presented for annotation are tiny, each containing a single object among the categories under consideration.[12] In the easy cases, the object to be identified is clear; in the difficult cases the images are hard to identify. As with RTE, we found disagreements in this task were largely due to this item difficulty. Even when the images are tiny or distorted, so that annotators disagree on what they refer to, they still refer to real-world objects that cannot be labelled in multiple ways.

(a) *cat* (b) *airplane* (c) *deer*

Figure 6: Some images in CIFAR with less than 0.3 observed agreement

10. The connection between disagreement and difficult can also be found in Beigman and Beigman-Klebanov (2009); Reidsma and Carletta (2008), and Beigman-Klebanov and Beigman (2014), among other previous studies.
11. It is interesting to note that the randomly selected perfect agreement instances are all $True$ according to the gold standard. Statistics show that the observed agreement for the gold $True$ class is on average higher than that of the $False$; annotators found it easier to identify entailment than to identify non-entailment.
12. Some images contain people or scenery, neither of which is a category in CIFAR.

Table 3: 10 randomly selected polarizing items in RTE

	Premise	Hypothesis	Gold	Observed Agreement
1	MSF was unnerved by a Taliban accusation that its members were spying for the U.S.	Taliban spies on U.S.	False	0.44
2	Al Thawra added, “Lahoud is well aware of that and realizes that Israeli challenges have never stopped for one moment; and that escalation will not hamper him from undertaking his national duties, relying on the support of all of Lebanon with all of its factions, as well as on the full support of Syria in order to achieve his national tasks and deliver on his commitments."	The newspaper added that regardless of the Israeli challenges, Lahoud would still be able to deliver on his duties, supported by Syria and a united Lebanon.	True	0.44
3	Al-Koshah’s events had surfaced after the British “Sunday Telegraph"; newspaper published last October 25 an article in which it accused Egyptian police of “;crucifying and raping Copts"	Was events Al Kusheh case emerged after the publication; newspaper Sunday Telegraph; “the British on 25 last October writes an accused in which the Egyptian police"; with steel Copts and rape of their Families	False	0.44
4	Seiler was reported missing March 27 and was found four days later in a marsh near her campus apartment.	Abducted Audrey Seiler found four days after missing.	True	0.46
5	The Bugbear virus infects computers running the Windows operating system and an unpatched version of Internet Explorer 5.5.	Virus infects thousands of computers.	False	0.46
6	Britney Spears is getting hitched for the second time this year - this time to a professional dancer father whose girlfriend of three years is pregnant.	Britney Spears is pregnant	False	0.46
7	In turn, the Editor-in-Chief of Al Jumhoria Newspaper was appointed Ambassador of Iraq to India.	Al Jumhoria is the Iraqi Ambassador to India.	False	0.46
8	Two Western citizens, one of whom is British, three policemen and two kidnappers were wounded in the attack that ended in the arrest of 13 kidnappers.	Wounded nationals statement one British and three police and in the attack which ended the arrest 13	False	0.46
9	German Chancellor Gerhard Schroeder accused U.K. Prime Minister Tony Blair and Italian Prime Minister Silvio Berlusconi of allying with European conservative parties in a “blockade" of the German and French-backed candidate, Belgian Prime Minister Guy Verhofstadt.	Schroeder doesn’t support Vershoftstadt as a candidate.	False	0.46
10	Johnston is the seventh person to be killed in sectarian violence this year in Northern Ireland where the outlawed IRA is fighting to end British rule in the province.	IRA killed Johnston.	False	0.46

Table 4: 10 randomly selected high agreement items in RTE

	Premise	Hypothesis	Gold	Observed Agreement
1	Iraq has been under a stringent economic embargo since its August 1990 invasion of Kuwait and relief workers are increasingly concerned about the health of its population.	An embargo was imposed on Iraq in 1990.	True	1.0
2	The three-day G8 summit will take place in Scotland.	The G8 summit will last three days.	True	1.0
3	Kidnappings in Argentina have increased more than fivefold in the last two years, official figures show.	Argentina sees upsurge in kidnappings.	True	1.0
4	But even in light of this unparalleled decline, the SPD's result in the June 13 European elections is of a qualitatively different character.	European elections took place on June 13.	True	1.0
5	A federal jury needed just four hours to return a death sentence against Chadrick Fulks, who pleaded guilty to kidnapping and carjacking resulting in the death of an Horry County woman.	Chadrick Fulks gets the death penalty	True	1.0
6	The G8 summit, held June 8-10, brought together leaders of the world's major industrial democracies, including Canada, France, Germany, Italy, Japan, Russia, United Kingdom, European Union and United States.	Canada, France, Germany, Italy, Japan, Russia, United Kingdom and European Union participated in the G8 summit.	True	1.0
7	Crude Oil Prices Slump	Oil prices drop	True	1.0
8	Last July, a 12-year-old boy in Nagasaki - a city just north of Sasebo - was accused of kidnapping, molesting and killing a 4-year-old by shoving him off the roof of a car garage.	Last year a 12-year-old boy in Nagasaki was accused of murdering a four-year-old boy by pushing him off a roof.	True	1.0
9	Shrek 2 retained the top spot with $92.2 million over the long Memorial Day weekend, fending off the global-catastrophe tale 'The Day After Tomorrow' which debuted with $86 million, according to studio estimates Monday.	Shrek 2 retained the top spot with $92.2 million over the long Memorial Day weekend, fending off the global-catastrophe tale 'The Day After Tomorrow' which debuted with $86 million, according to studio estimates Monday.	True	1.0
10	Ghazi Yawar, a Sunni Muslim who lived for years in Saudi Arabia, has been picked as president of Iraq after the favored U.S. choice, Adnan Pachachi, declined to take the job.	Yawer is a Sunni Muslim.	True	1.0

(a) *horse* (b) *ship* (c) *dog*

Figure 7: Some images in CIFAR with perfect observed agreement

2.8.5 Subjectivity

To conclude, we briefly discuss here another source of disagreement common in certain NLP tasks: **subjectivity**. As discussed in Section 2.7, in tasks such as offensive language identification, annotators may disagree on whether a segment of text is offensive or not, not because of interface issues, an overlap between categories, or because they are not paying sufficient attention, nor because the items are difficult to understand, but because they have different views on whether a segment counts as offensive or not (Akhtar et al., 2019). For instance, the Sexism dataset from Waseem (2016) consists of tweets such as (4) (reported by Akhtar et al. (2019)) classified by expert annotators and the crowd as either sexist or not.

(4) @ XXX uh... did you watch the video? one of the women talked about how it's assumed she's angry because she's latina.

Very low intercoder agreement is observed for such items, which are also flagged as being polarized by methods such as those proposed by Akhtar et al. (2019). This is because people have different subjective views on what counts as sexist or not.

In this study we focused on "objective" judgments, and thus none of the datasets studied in this paper cover this type of disagreement; we mention it here only for completeness. But it should be clear that such cases present the most serious challenge to the very idea of the "gold label," as any single label assigned to items such as (4) would be purely arbitrary.

3. Approaches to Learning from Disagreement

Current methods for learning from crowd annotations can be divided in four broad categories, summarized in Table 5:

1. Methods that automatically aggregate crowd annotations into (typically, one) single label for each instance. Most, although not all, of these models operate under the assumption that a single, objective ("gold") truth exists for every instance and aim to produce the best estimate of this truth without requiring manual adjudication, and ideally not even expert judgments. (The term **silver** truth is sometimes used for these automatically aggregated labels.)

2. Methods that still assume that a gold label exists for every item, but relax the assumption that this true label is always recoverable and use information about disagreement to either eliminate (**filter**) items whose gold label does not appear to be recoverable due to excessive disagreement among coders (**hard** items) or to **weigh** them.

3. Methods that can be used to learn a classifier directly from the crowd annotations, possibly via a (typically probabilistic) distribution that assigns a score to each label (**soft label**) computed from the crowd annotations using, e.g., softmax.

4. Methods that involve training a classifier using a combination of hard labels and soft labels extracted from crowd annotations. These methods use gold labels or estimate ground truths in training, but supplement these with information from the crowd annotations, e.g., to weigh an item by its estimated difficulty or the ability of its annotators.

Table 5: Taxonomy of learning from disagreement. Filter: whether items are filtered out; hard labels (single ground truth); soft labels: learning from multiple annotations.

Category	Example approache	Filter	Hard	Soft
Aggregation of coder judgements	Dawid & Skene, CrowdTruth		✓	
Filtering hard items	Reidsma & op den Akker, Beigman-Klebanov et al.	✓	✓	
Learning directly from crowd annotations	DLC, Soft loss, CrowdTruth			✓
Augmenting hard labels w/ disagreements	PEWE, Multi-task learning		✓	✓

The rest of this section reviews the research that has been carried out in each of these directions. For each of these categories, we will briefly discuss a few key research projects and papers, with emphasis on the details of a method and its underlying assumption about truth. The evaluation criterion for a model will also be noted. In Section 5, we will list the state-of-the-art and commonly used methods that we selected for in-depth analysis for each category, providing details about their implementation.

3.1 Aggregating Coder Judgments

The simplest way to automatically aggregate a multiplicity of annotations is **majority voting** (MV). Using this method, the estimated label for a given item is simply the label which receives the most annotations. Majority voting is simple to understand and implement, and it can produce good results when the annotators are in agreement with each other and with the experts, but it makes one key assumption that does not always apply: that all annotators are equally adept at the task. Further, majority voting does not take the level of difficulty of an instance into account in producing an aggregated label. These limitations are well known, and much research has aimed at addressing them.

Probabilistic aggregation methods Possibly the first and definitely the most widely used method attempting to address the limitations of majority voting was proposed by Dawid and Skene (1979). Their approach (henceforth, D&S) estimates the posterior probability of a label l_i for instance i conditioned on the observed label y, the actual label z_i, the prevalence of the labels π, and the probability $\theta_{j,k,k'}$ that annotator j assigns label k' to an item given its actual label is k (this latter probability is estimated for each coder from his/her annotations):

$$p(l_i|y, \theta, \pi) \propto p(z_i|\pi)p(y|l_i, \theta)$$

Numerous other probabilistic models for estimating ground truth have been proposed since D&S. Some of the most widely used include Smyth et al. (1994); Carpenter (2008); Whitehill et al. (2009); Hovy et al. (2013); Kamar et al. (2015); Moreno et al. (2015); Felt et al. (2015) and Li et al. (2019) (see Paun et al. (2018) for an overview and comparison of some of these models for NLP applications).[13] A model that has proven effective in many NLP applications

13. A great many other surveys of aggregation methods exist. Among these, we will only briefly mention here those by Zhang et al. (2016) and Zheng et al. (2017) to explain how this paper positions itself in comparison. Zhang et al. (2016) analyzed the basic concepts of label quality, outlined major research outcomes in methods for harnessing crowdsourced labels, and summarized some of available crowdsourcing datasets and tools. They also compared the performance of several ground-truth inference algorithms, noting that D&S is a standard method for label aggregation across several datasets, although it is out-

is the simpler MACE model by Hovy et al. (2013), in which the θ parameter of D&S is replaced by a parameter S_{ij} specifying the probability that coder j is spamming on i. Some of the models, such as Carpenter (2008); Whitehill et al. (2009) and Kamar et al. (2015), also model **item difficulty** (see below).

A non-probabilistic approach to aggregation particularly motivated by the intuition that disagreements are informative and not making the assumption that a gold truth exists was developed within the CrowdTruth project (Aroyo and Welty, 2015; Dumitrache et al., 2018c, 2019); we will discuss this approach next.

The CrowdTruth approach to aggregation The aim of the CrowdTruth project (Aroyo and Welty, 2015) was to investigate the hypothesis that "disagreement is signal, not noise." Research within this project led to the development of new metrics for assessing the quality of annotators, agreement on instances (a measure of item difficulty), and agreement on labels (Inel et al., 2014; Dumitrache et al., 2018c; Dumitrache, 2019), as well as revised versions of the standard precision/recall/F evaluation metrics discussed in Section 4. These methods were applied to relation extraction (Dumitrache, 2019), named entity recognition (Inel et al., 2014), and a variety of other tasks with both a closed and an open number of labels (Dumitrache, 2019).

Two versions of the metrics were proposed. In both versions, the computation of the metrics is based on two basic ingredients: a **worker vector** $w_{w,i}$ recording the answers of worker w on instance i and a **media unit vector** V_i summing up all the annotations of all the workers on instance i. These vectors are then used to compute quality metrics for workers, items (media units) and classes (annotations):

annotator quality: ($WQS(i)$)—the overall agreement of one worker with the other workers;

media unit quality: ($UQS(u)$)—the overall worker agreement on media unit u; and

annotation quality: ($AQS(a)$)—the overall agreement over an annotation, i.e., label, across all units in which it appears.

In the original version of the metrics as used, e.g., in Chapters 2 and 3 of Dumitrache's dissertation, these quality metrics were computed using cosine similarity to standards (e.g., the media quality—the extent to which an instance was a good example of a particular class—was computed by measuring the cosine similarity between that instance's media unit vector and the unit vector with a 1 for the class and 0 for all other classes). In version 2.0, all scores are mutually dependent on each other and are therefore computed through an iterative process.

In Dumitrache's work in particular, these quality metrics were then used to assign one or more classes to an instance i: every label whose score for i was higher than a certain (empirically established) threshold was considered a label for that instance.

performed on datasets with a high number of categories. Zheng et al. (2017) carried out an experimental survey of several ground-truth estimation techniques. They also recommend D&S, as it is robust across several tasks with very little overhead especially in terms of computational efficiency. The main difference between this study and surveys such as Zhang et al. (2016) and Zheng et al. (2017) is that in this paper, ground-truth inference algorithms are only one category of the experimented methods.

These metrics were shown to work better than MV in Dumitrache's work, but were not compared against other aggregation methods or other methods for using crowd annotations; we carry out this comparison in this survey.

Heuristic- and metric-based aggregation methods Many heuristic-based aggregation methods have been proposed (for a review see, e.g., Quoc Viet Hung et al. (2013); Sheshadri and Lease (2013), and Daniel et al. (2018)), but none of these have been shown to outperform D&S when the estimated ground truth is compared to gold labels.

Aggregation methods considered in this survey Dawid and Skene (1979) remains the most widely used probabilistic method for aggregating crowd annotations. As such, it is the main aggregation model analyzed in this survey. We also used majority voting as a baseline, as well as the two best-known aggregation methods proposed in the literature, MACE and the CrowdTruth approach to aggregation. We provide more details on the aggregation methods tested in this paper in Section 5.

3.2 Filtering Hard Items

Manual adjudication and automatic aggregation both result in a single gold or estimated (silver) label for all instances in a dataset that can then be fed to a supervised classifier. Models trained using such data are usually also evaluated assuming that a single label exists for each instance in the data. In this traditional approach, even substantial disagreement on a training/testing instance does not result in the removal of that instance. Several researchers, however, have argued that information about disagreement should be used to filter the dataset; items on which there is substantial disagreement should not be used to train or evaluate models.

For instance, Reidsma and op den Akker (2008) consider inter-annotator disagreement to be an indicator of how easy or difficult an item is. They consider two ways of using disagreement to improve the performance of a classifier. The first method is to filter the data by training on the high agreement subset of the data only, i.e., treating the other items as noise. The second, softer approach is to train several classifiers, one for each annotator, and to build a "voting classifier" that makes a prediction when all the classifiers agree on the class label. Both methods have been shown to have high precision but low recall when evaluated on test data containing instances with varying levels of agreement.

A more systematic analysis of the effect of noisy items was carried out by Beigman-Klebanov and colleagues (see, e.g., Beigman Klebanov et al. (2008); Beigman-Klebanov and Beigman (2009); Beigman and Beigman-Klebanov (2009), and Beigman-Klebanov and Beigman (2014)). Beigman-Klebanov et al. argue that low agreement on an item suggests it is not a good example for the phenomenon at hand, as it introduces noise in a model at training time and cannot be fairly assessed at test time; it should therefore be **filtered** from the training and test data, or at the very least separated from the high agreement (easy) cases and trained on and evaluated separately. Beigman-Klebanov and Beigman (2009) proposed a model of "hardness" that can be used to carry out this filtering or separation, but did not test this model. Beigman-Klebanov and Beigman (2014) proposed a simpler model based on a categorization of items ranging from "very easy" to "very hard" depending

on the extent of the disagreement and compared the effect of selecting subsets of items at training and test time.

One important issue concerning "hardness" is the observation by Reidsma and Carletta (2008) that not all disagreements are equally problematic for a machine learning algorithm. Disagreements are unproblematic as long as they can be viewed as random noise; they become an issue when they reveal the existence of different annotator biases, which, according to Reidsma and Carletta, are revealed by the appearance of patterns of disagreement. A proper model of hardness ought to capture this finding.

Several models of hardness have been explored in the literature in addition to Klebanov et al.'s. Arguably, the theoretically most developed approaches are the models of item difficulty incorporated in popular probabilistic aggregation models such as Carpenter's (Carpenter, 2008) and Whitehill's GLAD model (Whitehill et al., 2009). We tested item difficulty based on high inter-annotator disagreement as in Reidsma and op den Akker (2008) and computed using Whitehill's GLAD model, which is the most widely used.

3.3 Learning a Classifier Directly from Crowd Annotations

The third category of approaches to learning from the crowd consists of methods that do not make the assumption that a single gold label exists or is recoverable (and thus do not aim to identify a silver label, although they may weight labels according to various factors) and/or aim to capture the intuition that the distribution of labels produced by the crowd provides useful information (and thus do not attempt to filter items on which substantial disagreement is observed). Such methods therefore attempt to learn a model directly from the crowd's annotations. Broadly speaking, one can find three varieties of this class of methods in the literature:

(1) methods that treat each annotation as a separate learning instance;

(2) methods that aggregate the annotations into a probabilistic distribution (**soft label**), then learn directly from that distribution using a soft loss function;

(3) methods that estimate a probabilistic soft label as above jointly with learning a classifier.

We next discuss paradigmatic approaches for each of these sub-categories.

Repeated labelling The first type of approach is exemplified by one of the best-known proposals in this area: Sheng et al. (2008)'s **repeated-labelling** approach (SREL). Sheng et al. (2008) proposed that for each instance x, replicas of x are created for each unique label j assigned by the crowd to x. A distinct replica may be created for each annotation, or a replica may be created for each label, but weighted appropriately (e.g., it can receive a weight of $\frac{1}{|x^j|}$ or $|x^j|$, where $|x^j|$ is the number of annotations of x with label j). This approach is the sole soft-labelling approach tested by Jamison and Gurevych (2015).

Soft loss functions A second but equally intuitive way to train directly from the annotations is to use the probability distributions of item labels as soft targets in a loss function that can be used with such labels, such as cross-entropy (henceforth CE) or mean square error (MSE). Recently, Peterson et al. (2019) have provided evidence of the benefits obtained

using such soft labels applied to training a computer vision model for image classification. Peterson et al. (2019) have argued that the cross entropy between the soft label and the probabilistic distribution predicted by the model is the optimal loss function when the goal is to generalize well to unseen data. They showed that training using soft loss outperforms training on the single gold. Consider the standard loss function:

$$min \sum_{i=1}^{n} L(f_\theta, x_i, y_i), \tag{5}$$

where L is the loss for a model with parameters θ and the objective is to minimize L with respect to some observed data $\{x_i, y_i\}_{i=1}^{n}$, such that the function f_θ generalizes well, i.e., minimizes the expected loss over unobserved labels given previously unseen input data $\{x_j\}_{j=1}^{m}$ drawn from the same data distribution:

$$\sum_{i=1}^{m} \sum_{c} L(f_\theta, x_i, y_j = c) p(y_j = c | x_j). \tag{6}$$

Uma et al. (2020) further explored this approach by applying it to datasets from NLP as well as CV, using different types of characteristics, and considering alternative probability-comparing functions such as the Kullback-Leibler divergence. We discuss these options in more detail in Section 5.2.

Training using soft labels also bears similarities to using self-supervised loss functions in the semi-supervised learning literature. For example, in temporal ensembling (Laine and Aila, 2016) a secondary distance loss on pseudo-labeled samples (e.g., squared distance) is used to guide learning, albeit with a different goal.

Learning from crowds Raykar et al. (2010) pioneered the **learning from crowds** approach of carrying out aggregation while jointly training a model. Building on Dawid and Skene (1979), Raykar et al. (2010) used the expectation maximization algorithm to jointly learn an estimated gold label, annotator reliability, and a classifier to predict whether a suspicious region on a medical image from an X-ray, CT scan, or MRI was malignant or benign. The annotations used in their experiments were provided by expert radiologists, and the model learned was a logistic classifier, but they argue that the model can be used for any classifier and in a multi-class setting.

An extension of this approach to a deep learning setting was later proposed by, for example, Albarqouni et al. (2016), who developed a multi-scale CNN, *AggNet*, to handle data aggregation directly as part of the learning process via an additional crowdsourcing layer. Albarqouni et al. (2016) also exemplified their method using histology images in a binary classification setting. Guan et al. (2017), too, propose a neural network model for learning from medical experts (in this case, learning diabetic retinopathy severity on a 5-point scale). However, their model learns from multiple annotators (also experts) by modeling them individually with a shared net that produces unique outputs for each expert, and in addition learns averaging weights for combining their modeled predictions (Guan et al., 2017).

Most recently, Rodrigues and Pereira (2018) proposed a similar approach to Guan et al. (2017) that they called **deep learning from crowds** (henceforth, DLC). DLC not only

learns to combine the votes of multiple annotators, but also captures and corrects their biases while remaining computationally less complex than Guan et al. (2017). Deep learning from crowds was shown to work for binary classification, multiclass classification, regression, and structured prediction problems, both in CV and NLP. In their paper, Rodrigues and Pereira (2018) show that their model outperforms existing models including Guan et al. (2017) when evaluated against a gold truth (see below). We chose DLC as the paradigmatic model for this sub-category in part because it had the best performance when our experiments started, and in part because it has been applied both to CV and NLP problems.

3.4 Using both Hard Labels and Information about Disagreements

Finally, a range of methods have been proposed that assume the existence of a gold or silver hard label for each item but also recognize that uncertainty is a real possibility, and thus models may benefit from leveraging information from crowd annotations. Such methods can be further subdivided into:

(1) methods that use the crowd annotations to estimate the uncertainty on the label—and use this estimate to weight the loss associated to an item—and

(2) methods that jointly learn from the hard labels and the additional information (soft labels or item difficulty).

Plank et al. One of the best known proposals regarding learning from disagreements in NLP, the method by Plank et al. (2014a) (which we will refer to as **pewe**) falls under the first sub-category.

Plank et al. computed the extent of confusion on a label from inter-annotator agreement between two expert annotators on a small sample, and used that overall degree of confusion between labels to weight items while learning a part-of-speech (POS) tagging model from gold labels. They tested two different ways to quantify this label uncertainty, F1-score and tag confusion probability, finding that tag confusion probability outperformed F1 score (Plank et al., 2014a).

Sharmanska et al. A number of alternative approaches also using label confusion or item difficulty to weight the hard label have been proposed. For instance, Sharmanska et al. (2016) used inter-annotator agreement to discriminate between easy and difficult examples, but like Plank et al. (2014a), they integrated this information into their classifier as a measure of confidence in the usefulness of the data instance instead of using it to filter the instance. They did this using a model based on Gaussian processes in which "annotation ambiguities" informed the likelihood function of the classifier regarding whether the influence of a given item should be retained, reduced, or ignored. Their work was not concerned with the availability or lack of ground truth; rather, they focused on instance weighting and attempted to use disagreement to inform how much importance the learner should ascribe to each instance in the data (Sharmanska et al., 2016).

Jointly learning from gold and disagreements Lalor et al. (2018) proposed training algorithms in which both gold labels and soft labels were used at different times—either using gold labels for one epoch and soft labels for the next, or training using gold labels and then fine-tuning using soft labels.

More recently, Fornaciari et al. (2021) exploited crowd disagreement in a multi-task learning (MTL) setting. They trained models for two common NLP tasks—POS-tagging and stem identification—which jointly learn to classify the hard labels (i.e., gold-standard classes) and the soft labels (i.e., the coders' annotations), represented as a probability distribution. They used cross-entropy loss for the gold classification task but Kullback-Leibler divergence to estimate the error between the predicted and the target probability distribution.

Fornaciari et al. compared models trained in this MTL setting with single-task learning (STL) models, which only learn the gold classification task without exploiting disagreement. Their results show that the MTL models consistently outperform the STL ones. They interpreted the contribution of the secondary task in terms of regularization: The back-propagation of the loss tends to penalize errors more for the instances where there is a peak of probability, i.e., where the coders agree more. In contrast, there is less of a penalty in instances with smoother distributions, i.e., where the coders disagree.

3.5 Coming to Soft Labels from Another Direction: Learning from Noisy Labels and Distillation

As mentioned in the Introduction, a very active line of research in AI in general and computer vision/NLP in particular is the study of methods to learn from noisy labels above and beyond the noise due to disagreements between annotators (Mnih and Hinton, 2012; Northcutt et al., 2019). A line of work particularly relevant to this paper focuses on methods that *introduce* a measure of noise in the labels in order in order to improve generalization. Among such methods, the best known is perhaps **distillation**, proposed by Hinton et al. (2015). Distillation is a technique for transferring knowledge from a more complex, "teacher' model to a smaller, "pupil" model. One of the key ideas is that distillation works best when the student learns from the entire probability distribution assigned to an item by the teacher instead of from a single label. Although there is a connection between learning from soft labels containing disagreements originating from human judgments and learning from (naturally or artificially) noisy labels in general (a connection also highlighted by, e.g., Peterson et al. (2019), who compare their models for learning from disagreement with models for learning via distillation), methods in which the soft label is generated from models are outside the scope of this study, which focuses on learning from naturally generated soft labels.

4. Evaluation

As we have seen, there is by now an extensive literature on learning from multiple annotations, possibly in disagreement. Much, although not all, of this work is motivated by empirical findings such as those discussed in Sections 1 and 2 suggesting that gold labels are only an idealization, at least for cognitive tasks. Yet, much less research has been devoted to the the question of how to evaluate models in such circumstances, especially when the "true" label is not known (i.e., our **RQ1**).

Two forms of evaluation have been used in the literature on learning from data containing multiple annotations of the same item. **Hard** evaluation metrics such as accuracy or F1 are traditionally used when it is assumed that a true label exists notwithstanding the

disagreement between annotators. More recently, however, evidence such as that presented in Section 2 has led researchers to question the validity of evaluating models trained with data collected without assuming a gold label against test data with gold labels (e.g., using accuracy). Therefore, in this survey, all models were evaluated using not only traditional, "hard" evaluation metrics, but also **soft** evaluation metrics which do not involve a gold label, some already used in the literature but not yet established, some novel. In this section we discuss the options we considered.

4.1 Hard Evaluation Metrics

Two types of metrics have been used for hard evaluation against gold labels, measuring respectively:

1. ***How well the model predicts gold labels when all items are treated equally***: This is the traditional "hard" way of measuring model performance, and it is interesting to note how many proposals for learning from disagreement were evaluated this way, even among those arguing that assuming a gold label exists is too strong an idealization (Sheng et al., 2008; Plank et al., 2014a; Alonso et al., 2015; Sharmanska et al., 2016; Rodrigues and Pereira, 2018). The most frequently used hard measures include percentage **accuracy** and class-weighted **F1** with respect to the gold labels; we used both in this study.

2. ***How well the model captures truth when items are weighed depending on disagreement***: An alternative approach to evaluating models using disagreement as signal is the crowd-truth weighted f-measure (henceforth, **CT F1**) (Dumitrache et al., 2018c). This metric still relies on a "hard" label, but does not give the same weight to all items. The intuition is that items on which there is a lot of disagreement ("hard" items) should be weighted less than "easy" items; hence an inverse-confusion score (a "sentence relation score," which we will call here the "item relation score" (*irs*) for the sake of generality) is used to weight the standard precision and recall scores, resulting in the weighted precision, $P^{'}$, and weighted recall, $R^{'}$, defined as:

$$P^{'} = \frac{\sum_i irs(i) \ * \ tp(i)}{\sum_i irs(i) \ * \ tp(i) + (1 - irs(i)) \ * \ fp(i)} \tag{7}$$

$$R^{'} = \frac{\sum_i irs(i) \ * \ tp(i)}{\sum_i irs(i) \ * \ tp(i) + irs(i) \ * \ fn(i)} \tag{8}$$

   The weighted f-measure, CT F1, is then defined as usual as the harmonic mean of the weighted precision, $P^{'}$, and weighted recall, $R^{'}$ (Dumitrache et al., 2018c).

   $irs(i)$ is defined as the cosine similarity between the item vector and the unit vector for the label under consideration; a higher irs implies that a majority of annotators agreed with the gold labelling. Formally, $irs(i) = cos(\boldsymbol{V}_i, \hat{r})$ where $\boldsymbol{V}_i$ is the media unit vector discussed in Section 3.1 (the item's label distribution) and $\hat{r}$ is a vector whose dimension is the number of labels, with 0 values for all components except for the component corresponding to relation r.

4.2 Soft Evaluation Metrics

No generally accepted form of soft evaluation exists if the existence of a gold label is not assumed. Therefore, we considered a variety of approaches, measuring:

1. ***How similar the distribution of labels assigned by the model to an item is to the distribution of judgments produced by the annotators for that item***: This type of evaluation captures the ability of the model to learn the probabilities of each label relative to the others for a given instance. The underlying assumption is that the item label distribution produced by the annotators is representative of the implicit ambiguity of each item. Given a set of inputs, $\mathbf{x} = \{x_i\}_i^m$, if we define $p_{hum}(x_i)$ to be the probability distribution of the crowd annotations over the set of labels for that item and $p_\theta(x_i)$ as the probability distribution for that item produced by a model with parameters θ, we measured this similarity in two ways:

   - Peterson et al. (2019) proposed evaluating the trained models using **cross entropy** (CE) in order to capture how confident the model is in its top prediction compared to humans and the reasonableness of its distribution over alternative categories.
   $$\text{CE}(p_{hum}(\mathbf{x}), p_\theta(\mathbf{x})) = \sum_{i=1}^{m} p_{hum}(x_i)\ log\ p_\theta(x_i). \tag{9}$$

   - **Jensen-Shannon divergence** (JSD) (Lin, 1991) is a standard method for measuring the similarity between two probability distributions. It is based on the Kullback-Leibler divergence (Kullback and Leibler, 1951) (KL), but is symmetric and always has a finite value (see discussion below).
   The Jensen-Shannon Divergence between $p_{hum}(x_i)$ and $p_\theta(x_i)$ can be expressed in terms of KL divergence as follows:
   $$JSD(p_{hum}(x_i) \parallel p_\theta(x_i)) = \frac{1}{2} D_{KL}(p_{hum}(x_i) \parallel M) + \frac{1}{2} D_{KL}(p_\theta(x_i) \parallel M) \tag{10}$$
   where $M = \frac{p_{hum}(x_i) + p_\theta(x_i)}{2}$.
   $D_{KL}(p_{hum}(x_i) \parallel p_\theta(x_i))$ denotes the KL divergence between the two distributions and is computed as:
   $$D_{KL}(p_{hum}(x_i) \parallel p_\theta(x_i)) = p_{hum}(x_i)\ log\ \frac{p_{hum}(x_i)}{p_\theta(x_i)} \tag{11}$$
   Using Jensen-Shannon Divergence, the similarity can be expressed as:
   $$JSD(p_{hum}(\mathbf{x}), p_\theta(\mathbf{x})) = \sum_{i=1}^{m} JSD(p_{hum}(x_i) \parallel p_\theta(x_i)) \tag{12}$$

2. ***How well the model captures human uncertainty in its prediction***: An alternative approach to evaluating a model's ability to reproduce human judgments is to evaluate the model's ability to capture the disagreements among annotators in annotating the item as measured using **normalized entropy**. The assumption is that

the entropy of the annotators' distribution is a good measure of how confusing the annotators find the item. To measure the ability of a trained model θ to capture annotators' confusion, first, we compute on an item basis the normalized entropy of the probability distribution produced by the model, $H_{norm}(p_\theta(x_i))$, and the normalized entropy of the soft labels, $H_{norm}(p_{hum}(x_i))$, for each item i. We then compute the vectors of the entropy values over all the items, $\boldsymbol{H}_{norm\_hum}$ and $\boldsymbol{H}_{norm\_\theta}$. Finally, the model is evaluated using:

- the cosine similarity between the two vectors, which we call the **entropy similarity** metric:

$$sim(\boldsymbol{H}_{norm\_hum}, \boldsymbol{H}_{norm\_\theta}) = \frac{\boldsymbol{H}_{norm\_hum} \cdot \boldsymbol{H}_{norm\_\theta}}{||\boldsymbol{H}_{norm\_hum}||\,||\boldsymbol{H}_{norm\_\theta}||} \tag{13}$$

- Pearson (1896)'s correlation between the two vectors, which we call the **entropy correlation** metric:

$$corr(\boldsymbol{H}_{norm\_hum}, \boldsymbol{H}_{norm\_\theta}) \tag{14}$$

4.3 Comparisons of the Evaluation Metrics

In this section, we provide some intuition and explanations of our expectations for the evaluation metrics introduced in sections 4.1 and 4.2.

4.3.1 Accuracy vs F1 vs CT F1

As discussed in Section 4.1, three "hard evaluation" metrics can be used to capture the degree of correctness of the predictions of a model/method with respect to the expert provided target labels—accuracy, F1, and CT F1. The first two metrics do not take disagreement into account; the CT F1 metric on the other hand weights each item's contribution to the overall score by how confusing the annotators find that item. Hence, we had three expectations for the metrics. Firstly, we expected the relative rankings of the models to be largely similar using both accuracy and F1, except that because the F1 metric is class weighted, we expected that in datasets with class imbalances, the rank of the methods might be different, as it would be based on their performance on the majority class.

Second and third expectations about the hard metrics were discussed by Dumitrache et al. (2018c,a). To illustrate them, we'll use a simplified example. Consider a binary task with items belonging to either category 0 or 1, assuming that for the 4 items in the dataset, item relation scores (irs) are $[0.2, 0.2, 1, 0.8]$, respectively,[14] and the gold labels are $[1, 1, 1, 1]$. Then, consider three models m_1, m_2, and m_3. The predictions of these models and their F1 and CT F1 for class 1 on the hypothetical data subset are shown in Table 6. Model 1, m_1, is a perfect model, retrieving all the relevant items. m_2 retrieved the items with high irs (i.e., the items for which the annotators highly agree with the gold label). m_3 retrieved the low irs items. Two observations can be made from the table: (1) the margin between m_2 and m_1 is narrower for CT F1 than for F1 and (2) for m_2, the CT F1 score is higher than the F1 score. By de-emphasizing "hard" items, the CT F1 score allows models

14. Given the definition for irs, the more annotators agree with the gold interpretation, the higher the irs score.

Table 6: The predictions f1 and ct f1 for hypothetical models on a hypothetical binary task

Model id	Model predictions	f1	ct f1
m_1	[1, 1, 1, 1]	1.0	1.0
m_2	[0, 0, 1, 1]	0.5	0.53
m_3	[1, 1, 0, 0]	0.5	0.28

that perform well on "easy" items to achieve more competitive scores. This in line with the hypothesis of Dumitrache et al. (2018a): Making the assumption that, for all items, the gold label is always perfectly suited/related to the item underscores the models' performance. Dumitrache et al. (2018a) further state that low(er) f1 scores for models are caused by these "hard" items. If their hypothesis stands, we do not expect to see models behave like m_3, i.e., to have a negative differential between their ct f1 and f1 scores.

4.3.2 JSD vs Cross Entropy

As mentioned in Section 4.2, the jsd function is a standard way of measuring the difference between two distributions. In coding theory, kl divergence (also known as relative entropy) is often interpreted as the number of extra bits required to send messages using the distribution Q when the optimal distribution is P. In the machine learning and statistical literature, kl is often used to measure the amount of information lost when Q is proposed as an approximation of P (P typically represents the "true" distribution and Q a model's prediction). Mathematically, the relative entropy from Q to P (i.e., the relative entropy of Q with respect to P) is defined as follows:

$$D_{KL}(P \parallel Q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{15}$$

This implies that when $p(x) = q(x)$, $D_{KL}(P \parallel Q) = 0$.

cross entropy can also be interpreted using coding theory. While kl measures the number of *extra* bits per message, cross entropy is the *average* or *expected* number of bits needed to send messages using Q when the optimal distribution is P. Mathematically, the cross entropy of Q with respect to P is given as:

$$H(P, Q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{16}$$

Consequently, when $p(x) = q(x), H(P, Q) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$, which is the entropy of P—$H(P)$. In other words, the lower bound of $H_{KL}(P, Q)$ is not *necessarily* 0 but the entropy of P.[15] We can also re-formulate $D_{KL}(P \parallel Q)$ as follows:

$$H(P, Q) = D_{KL}(P \parallel Q) + H(P) \tag{17}$$

where $H(P)$ denotes the entropy of P.

15. In our context, the lower bound of $H_{KL}(P, Q)$ is only 0 when a given item belongs exclusively to a single class

The implications of these observations for the purposes of evaluating models by comparing their predicted distributions with the distribution produced by the annotators are as follows. Firstly, when a model's predictions are perfectly identical to the human distribution, the KL divergence (and the JSD) is always 0—i.e., both KL and JSD, but not CE, are lower bounded at 0. Secondly, we see from equations (15) and (16) that neither KL divergence nor CE have an upper bound. The Jensen-Shannon divergence function, however, is capped at $1 \ln 2$ for the log base e, or 1 if using the base 2 logarithm (Lin, 1991). This upper and lower boundedness makes JSD desirable as a metric, as the results are more easily comparable across datasets. Having both hard and soft metric scores on the same scale allows for the possible combination of both (for instance by taking the sum of half of each). This is the reason we chose JSD as one of our soft evaluation metrics.[16]

Finally, the fact that JSD scores are bounded within a small range, unlike CE and KL, also means that all the results are confined within a smaller range than the scores for KL divergence or CE. This might have implications for checking the significance of results for the same dataset across different models; a narrower range would mean the results might seem to converge to a point, making it difficult to tell at a glance which results significantly differed from each other. For this reason, we also kept the widely known and used unbounded CE metric as a soft evaluation metric. We did, however, expect the model rankings of both CE and JSD to be largely similar.

4.3.3 Entropy Similarity vs Entropy Correlation

We used (normalized) entropy to measure the degree of uncertainty in the prediction of the crowd or the model for any any given item. To compare the uncertainty of a model with respect to the uncertainty of the crowd, we used Pearson correlation (Pearson, 1896) and cosine similarity (see Section 4). While neither satisfies the triangle of inequality nor can be considered a metric in the mathematical sense, they both measure important relationships. With entropy correlation, we can measure the linear relationship between the two vectors. In other words, we can answer the question, "Is the model uncertainty high when the crowd uncertainty is high and low when crowd uncertainty is low?" With entropy similarity, bounded between [0, 1], we can get a sense of how similar the vector of model entropy is to the vector of crowd entropy.

It is worth noting that cosine similarity is correlated with correlation; the more similar vectors are, the higher their correlation. As such, we expected the models' ranking by entropy similarity and entropy correlation to be largely the same. It is also worth noting that since we used the normalized version of entropy, the results using these two "metrics" are comparable across datasets.[17]

4.3.4 Entropy Similarity/Entropy Correlation vs Cross Entropy/JSD

The reason for using both divergences (cross entropy and JSD) and the scalar measures entropy similarity/correlation (using entropy similarity/entropy correlation) can be illustrated by the following hypothetical example. Consider a scenario where $p_{hum}(y_i|x_i) =$

16. There has been some discussion in the literature about normalized cross entropy (Stemmer et al., 2002; Sohn, 2016), but this is not yet as widely accepted as JSD.
17. We refer to them as metrics for the purposes of this paper.

$[0.8, 0.2, 0.0, 0.0]$ for a given item i and two models m_1 and m_2 produces a $p_\theta(y_i|x_i)$ of $[0.6, 0.2, 0.2, 0.0]$ and $[0.0, 0.0, 0.8, 0.2]$, respectively. We can make two observations from this example. First, model m_1 agrees more with the crowd on where probability mass should be assigned to item i. Second, model m_2 totally disagrees with the crowd on which classes are valid for item i but has the same general level of uncertainty about its prediction as the crowd. With cross entropy and JSD we can capture the first type of similarity: The cross entropy and JSD scores for m_1 and m_2 are $(0.73, 0.33)$ and $(27.63, 1.0)$, respectively.[18] In contrast, with entropy similarity and entropy correlation we can capture the second type of similarity: $H_{norm}(p_{hum}(y_i|x_i)) = H_{norm}(p_{m_2}(y_i|x_i)) = 0.36$ while $H_{norm}(p_{m_2}(y_i|x_i)) = 0.69$.

5. An In-Depth Comparison of Methods for Learning from Disagreement

In previous sections we discussed several NLP and computer vision tasks that motivated the development of methods for taking information about disagreements into account (Section 2), surveyed the literature on learning from multiply annotated data (Section 3), and discussed how models trained using such methods can be assessed (Section 4). In order to analyze these methods in more depth, we carried out experiments with those that were the most representative, comparing them on the same tasks using some of the best-known datasets in this literature. In this section, we discuss the design of these experiments.

First of all, we discuss in Section 5.1 the models that we implemented for each of the six tasks studied in this research (see Section 2). Secondly, we list in Section 5.2 the prominent state-of-the-art methods for learning from disagreement that we identified as best exemplifying each of the classes of approach discussed in Section 3 and tested by using each of them to train the models in Section 5.1 for the six tasks in Section 2. For each task, we computed the significance of the difference in performance between each pair of trained models using bootstrap sampling, following Berg-Kirkpatrick et al. (2012) and Søgaard et al. (2014). Section 5.2 provides the necessary detail about these training methods and how they were applied to each task.

The results of this comparison are presented and discussed in the following sections. A summary of the results is presented in Section 6; these results are then analyzed on a task-by-task basis in Section 7, and an overall discussion is provided in Section 8.

5.1 The Models

Part-of-speech tagging For POS tagging, we implemented our own POS tagger, inspired by Plank et al. (2016) but deploying an attention layer over two kinds of input representations—character and word level—with each level of representation encoded using a separate RNN architecture. At the character level, each word was encoded as a sequence of characters (using a "sequence RNN"), and the final states for each sequence of characters were used as representations. To get word-level representations, each word was encoded by passing the word embeddings through a "context bi-RNN"; the word embeddings were initialized from pretrained Glove embeddings (Pennington et al., 2014). Each representation was passed through a separate attention mechanism (Yang et al., 2016). The final representation, a concatenation of these outputs, was passed through a FFN with one ReLU

18. We avoided infinite values by clipping the values using a small epsilon, 1e-12

hidden layer and an output layer with softmax activation so that the output of the model was the probabilities for each word belonging to each of the the 12 universal POS tags.

The model was always trained for 20 epochs using the Adam optimizer (Kingma and Ba, 2015) at a learning rate of 0.001 with the the model with best development F1 saved at each epoch. This best model was used to evaluate the test data.

Information status classification The model for this task was developed by combining elements from two architectures: the state-of-art coreference architecture and the state-of-art IS classification architecture. The state-of-the-art architecture for IS classification at the time we started these experiments (Hou, 2016) was developed for the ISNOTES corpus (Markert et al., 2012; Hou et al., 2013) and achieves a performance of 78.9% on that corpus. The state-of-art coreference resolution architecture at that time (Lee et al., 2018) includes a mention representation component. We developed our model by sorting the mentions using the algorithm outlined by Hou (2016) and a span representation similar to Lee et al. (2018) but including the non-syntactic features from Hou.

The IS model was trained for 10 epochs with training parameters set according to Lee et al. (2018). For each experiment, we chose the best model based on the development set.

Relation extraction For this task we fine-tuned a BERT sentence classifier (Devlin et al., 2019) for the binary (medical) relation extraction task. The predicted probability for a sentence was obtained by applying a softmax function over the classifier's 2D output.

The model was trained for 4 epochs using a 10-fold cross-validation at a learning rate of 2e-5. Although simple, this model performs much better than the original model by Dumitrache et al. (2018b), which only achieved a micro-averaged F1 of 0.638 on the same MRE dataset used here, whereas our model achieves a micro-averaged F1 of 0.847.

Recognizing textual entailment Jamison and Gurevych (2015) reproduced the basic RTE system described by Dagan et al. (2005), but this model is no longer state of the art, so we developed a new model. Given the small size of the dataset, the model had to be concise, with as few parameters as possible without sacrificing performance. Our system encodes the premise and hypothesis texts using BERT (Devlin et al., 2019) and concatenates the encoded pair. This concatenation is the sentence-pair representation and is passed through a feed-forward neural network with 3 ReLU activated hidden layers and an output layer. The predicted probability for each example pair is obtained by applying a softmax function over the outputs.

The model was trained for 20 epochs using 10-fold cross-validation and the Adam optimizer (Kingma and Ba, 2015) at a learning rate of 0.0001. When trained on gold labels, this model outperforms that in Jamison and Gurevych (2015). While the Jamison and Gurevych (2015) RTE achieved 51.3 micro F1, our model achieves 61.31 micro F1.

Image classification: LabelMe We replicated the model from Rodrigues and Pereira (2018) for this task. Rodrigues and Pereira (2018) encoded the images using pretrained CNN layers of the VGG-16 deep neural network (Simonyan et al., 2013). This encoding is passed into a feed-forward neural network layer with a ReLU activated hidden layer and a single output layer. Output probabilities are obtained by applying a softmax function to these outputs.

In our experiments, we randomly split the 10k images with crowd annotations into training and test data (8882 and 1118 images respectively) to allow for ground truth and probabilistic evaluation. We used 500 gold-labeled images from the dataset as our development set. Training was carried out for 50 epochs using the Adam optimizer (Kingma and Ba, 2015) at a learning rate of 0.001. The model with the best development F1 was always chosen as the model used for evaluation.

Image classification: CIFAR-10H The model trained for this task was the ResNet-34A model (He et al., 2016), a deep residual framework that is one of the best-performing systems for the CIFAR-10 image classification model. We used a publicly available Pytorch implementation of this ResNet model.[19]

We trained the model with for a total of 65 epochs divided into segments of of 50, 5, and 10, using a learning rate of 0.1 and decaying the learning rate by 0.0001 at the end of every segment. The model used for the evaluation phrase was the model with the best development dataset.

5.2 Learning from Disagreement: Approaches Tested

In order to compare methods for learning from disagreement, we trained each of the architectures discussed in the previous section using representative approaches from every category discussed in Section 3. In this section, we list the methods we tested, providing the essential details about how they were implemented or used. The methods are grouped according to the same categories as in Section 3.

5.2.1 Aggregating Coder Judgments

As discussed in Section 3, possibly the most common approach to using the labels produced by the crowd is to go through an aggregation step during which the labels used for learning are obtained either through manual adjudication or through automatic aggregation. This process is normally based on the assumption that each item belongs to a single category, but the result of this preliminary step may also be a graded ranking of the labels. We trained our models using data aggregated as follows:

1. **Gold Training** This is training using a single gold label per instance, usually obtained through the manual adjudication of annotations produced by at least two manual annotators. (All the datasets we employed provided gold labels, with the exception of PDIS, which only includes gold labels for the test data.)

2. **Majority Voting** This is training using for each instance the label chosen by the majority of coders.

3. **Dawid and Skene** This is training using, for each instance, a single label produced by choosing the label with the highest posterior probability as assigned by the Dawid and Skene (1979) algorithm, which infers a per-class model of an annotator's expertise. We used a publicly available implementation of the Dawid and Skene (1979) algorithm,[20]

19. The CIFAR-10 model is available at `https://github.com/KellerJordan/ResNet-PyTorch-CIFAR10`
20. The Dawid and Skene algorithm can be found at `https://github.com/sukrutrao/Fast-Dawid-Skene`

but unlike the original paper, which used random initialization, we obtained initial estimates of the ground truth using majority voting.

4. **MACE** As a probabilistic alternative to D&S, we also tested the MACE item-response model (Hovy et al., 2013), which only learns whether an annotator is spamming on a given instance. We used the freely available implementation of MACE provided by the authors.[21]

5. **CrowdTruth** As a final aggregation method, we tested the quasi-probabilistic approaches developed in the CrowdTruth project, which involves computing several "quality metrics"—annotator, item, and label—to assign a label to an instance (Dumitrache et al., 2018c).

   The quasi-probabilistic approach in Dumitrache et al. (2018a), Dumitrache et al. (2018d), and Dumitrache (2019) was used for Twitter event identification, news event extraction, sound interpretation, and medical relation extraction (MRE), the task experimented with in this paper. As discussed in Section 2, Dumitrache et al. (2018a) collected annotations for these tasks using disagreement-aware crowdsourcing, i.e., workers were presented with a multiple-choice task, selecting from the 14 possible relations all the suitable interpretations (labels) for each item.[22] Because the MRE dataset used here is the one used by Dumitrache et al. (2018a), and they provide the aggregated labels,[23] we used these provided aggregate labels for the CrowdTruth experiments on the MRE task.

   Dumitrache et al. (2018a) generated labels from MRE crowd annotations using the following instantiation of the general approach discussed in Section 3:

   (a) **worker vector,** $W_{s,i}$ For each worker i annotating a sentence s, the vector cell for each relation the worker selects is marked 1, whereas the vector cell for the relations not selected are marked 0.

   (b) **sentence vector,** V_s The sentence vector for each sentence is computed by summing up the worker/annotation vectors for all the workers. $V_s = \sum_i W_{s,i}$

   (c) **sentence-relation score,** $srs(s, r)$ The sentence-relation score is computed as the cosine similarity between the sentence vector and the unit vector for that relation, $srs(s, r) = cos(V_s, \hat{r})$, where $\hat{r}$ is a one-hot vector with size the number of relations, marked as 0 in all cells except the cell corresponding to the relation being computed. The idea is that the higher the sentence-relation score, the more clearly the relation is expressed in the sentence, and thus the lower the level of ambiguity.

21. The version of MACE used can be found at `https://github.com/dirkhovy/MACE`
22. We should point out that the annotations for the other datasets experimented with in this paper (POS, RTE, IC-LABELME, IC-CIFAR10H, and PDIS) were not collected using disagreement-aware crowdsourcing; annotators could only select one of the available categories for each item to be annotated, although different annotators could end up choosing different labels.
23. Dumitrache's aggregated labels can be found at `https://github.com/CrowdTruth/Medical-Relation-Extraction`

(d) **sentence-relation score threshold,** t This is a fixed value in the [0, 1] interval used to differentiate between negative and positive relations for a sentence. Given a sentence-relation score threshold t, sentences with an srs threshold above t were given a positive label, while sentences with srs below t received a negative label.

Given the srs score for a set of sentences, Dumitrache et al. (2018a) produced weighted labels for those sentences by (1) separating the sentences into negative and positive sets based on the srs threshold (which they chose after experimenting with several thresholds) and (2) re-scaling the labels of sentences in the negative categories in the [-1, 0] interval. They did this because the manifold model (Wang and Fan, 2014) used in the paper required labels in the [-1, 1] interval. For our binary MRE classifier, we marked sentences in the negative set 0 and sentences in the positive set 1. We also experimented with using corresponding weighted labels (using the srs score as weights for training) but found that this led to a slight decrease in accuracy and F1; we therefore report the training on the unweighted binary labels.

Extracting a single crowd ground truth using the methodology discussed above (i.e., computing the srs score and a 0.5 threshold) is equivalent to majority voting, as the label with the most annotations will still be selected as the preferred label. We thus adapted the CrowdTruth methodology to a multi-class, multi-label scenario by using a vector with as many components as the number of labels, where the components were the srs scores of the corresponding labels. (A similar approach was used by Dumitrache et al. (2019) to adapt the methods to a multi-class setting.) For this reason, we consider the CrowdTruth approach for other tasks apart from MRE to be a *soft-label* method.

5.2.2 Filtering Hard Items

The second group of methods examined uses information about disagreements to *exclude* or at least *weight* instances.[24] We tested the following methods:

1. **Agreement Filtering** This involves training using an aggregated label after filtering away examples with low observed agreement (Artstein and Poesio, 2008). It was proposed by Beigman and Beigman-Klebanov (2009), but there was no specific recommendation as to what the agreement cut-off ought to be. Jamison and Gurevych (2015) tested two heuristically chosen thresholds for each task: low agreement and high agreement. In our experiments with this approach, we tried several cut-offs, and the results were the same: a decline in performance for all tasks except IC-LABELME image classification. In the end, we reported results obtained by filtering items with observed agreement below the average observed agreement for that dataset (which differs from task to task, as in Jamison and Gurevych's work).

   The formula for computing observed agreement was the same as in Artstein and Poesio (2008). Given a set of items I indexed by i, a set of categories K indexed by k, and a set of coders C indexed by c, the observed agreement for each item, agr_i, is given as:

24. Instance weighting can also be categorized in the third category, "Learning directly from the crowd annotations."

$$agr_i = \frac{1}{c_i(c_i-1)} \sum_{k=1}^{K} n_{ik}(n_{ik}-1) \tag{18}$$

where n_{ik} is the number of times item i is classified as category k. This formula was designed under the assumption that the C was the same for each item, which does not hold for three of the four datasets used here. To accommodate this, we adjusted c to mean the number of coders annotating the given item.

$$agr_i = \frac{1}{\binom{c_i}{2}} \sum_{k=1}^{K} \binom{n_{ik}}{2} = \frac{1}{c_i(c_i-1)} \sum_{k=1}^{K} n_{ik}(n_{ik}-1) \tag{19}$$

2. **Weighting by observed agreement** We also tested a soft version of filtering that involves weighting items by their degree of item difficulty instead of removing them. The idea is to weight difficult items less, so that the model learns to pay less attention to those items and does not overfit on items for which the labels are difficult/ambiguous.

   We tried two versions of this approach. In the first version, the loss of each item was weighted by the observed agreement of that item. Using this method and learning using MV as the aggregated label, has the effect of possibly down-weighting items on which majority voting differs from the gold interpretation. No previous references were found for this model, and this work is possibly the first use of this observed-agreement weighting method.

3. **Weighting by item difficulty** A second version of the weighting used the inverse difficulty predictions generated by Whitehill et al. (2009)'s GLAD (generative model of labels, abilities, and difficulties) aggregation model. The model uses a maximum likelihood algorithm to simultaneously infer annotator expertise, image difficulty, and the most probable label; it was tested by Whitehill et al. for binary image classification tasks ("male" vs "female" image categorization and *"Duchenne"* or *"non-Duchenne"* smile image categorization).

   We implemented this model and used our implementation to make item predictions for the binary classification tasks RTE, MRE, and PDIS.[25] During training, we weighted the loss for each item by the the item's probability of correctness, an estimate that takes image difficulty and labeler quality into account as in Whitehill et al. (2009)'s model.

5.2.3 Learning Directly from the Crowd Annotations

The methods grouped in this category in Section 3 seek to train a model directly from the annotations provided by the workers, i.e., without first going through an aggregation step. One point worth emphasizing is that in most cases, these models were originally evaluated assuming only a single ground truth. As far as we are aware, this survey is the first time all

25. There is little literature generalizing item difficulty models to the multi-label classification case—see Paun et al. (2021).

of these approaches have been evaluated using both "hard" and "soft" evaluation metrics, as discussed in Section 4.

1. **Sheng et al.'s repeated labelling** Sheng et al. (2008) proposed training a model directly from multiple annotations by presenting each input-annotation pair to the network as if it were a separate item. This was done as specified for 4 of the 6 tasks: POS, IC-LABELME, RTE, and MRE.

   Because PDIS has over 90k markables, each annotated 7 times on average, and CIFAR10H has 10k items annotated an average of 51 times, and because the classification model is quite complex, treating each annotation as a separate item for these tasks becomes unfeasible. Thus for these datasets, the models were fed each unique label only once, but the loss for each label was weighted by the number of times that label was chosen.

2. **Soft loss functions** As discussed in Section 3, Peterson et al. (2019) recently argued for training using the probabilistic distribution of labels obtained from crowd annotations (aka probabilistic soft labels) as a target.

   Using the human label distribution $p_{hum}(y|x)$—rather than the hard label from some consensus (adjudicated gold labels or aggregated labels)—and a negative log-likelihood, the loss proposed by Peterson et al. (2019) reduces to the loss function:

   $$-\sum_{i=1}^{n}\sum_{c} p_{hum}(y_i|x_i)\ \log p_\theta(y_i = c|x_i) \tag{20}$$

   where $p_\theta(y|x)$ is obtained by applying a probability function (softmax) over the logits produced by the classifier. This combination of target soft labels with a probability-comparing loss function is what we call the **soft loss function** approach.

   Uma et al. (2020) tested this hypothesis on 3 multi-class tasks with varying annotator and annotation characteristics: POS tagging, IC-LABELME image classification, and IC-CIFAR10H image classification. They then compared it to a number of existing models for learning from crowds. In this paper we report a more extensive investigation. We also tested the model on three binary classification tasks with varying characteristics. In addition to the cross-entropy (CE) loss function proposed by Peterson et al. (2019) and tested by Uma et al. (2020), we also tested other loss functions that can be used to minimize the difference between probability distributions. In particular, we tested using as loss functions mean-squared error (MSE) and Kullback-Leibler distance (Kullback and Leibler, 1951) (KL).

   Uma et al. (2020) also extended Peterson et al. (2019)'s empirical validation of the hypothesis in another direction. They hypothesized that although the standard normalization function works well for datasets like IC-CIFAR10H annotated by high-accuracy annotators and having a positive coder:label ratio, for datasets annotated by low-accuracy annotators that have a negative coder:label ratio, an exponential function (the softmax function) would be more appropriate. We continued this exploration,

extending the idea further to experimentally compare probabilistic soft labels estimated using standard normalization and softmax to the probabilistic soft labels that are the probabilistic posterior of the aggregation methods discussed in 5.2.1.

3. **(Deep) learning from crowds (Rodrigues and Pereira, 2018)** This approach involves adding a bottleneck layer, called a "crowd layer," after the output layer during training, so that the model learns the annotator matrix and thus how much weight to assign to each label. If the output of a neural network model is denoted by $\boldsymbol{\sigma}$, such that σ_c corresponds to the score assigned by the model to the input instance belonging to class c, then the activation of the crowd layer for each annotator, r, is defined as $\boldsymbol{a}^r = f_r(\boldsymbol{\sigma})$, where f_r is an annotator-specific function (Rodrigues and Pereira, 2018). The output of the crowd layer is simply the softmax of the logits (Rodrigues and Pereira, 2018). The generalized architecture for this is illustrated in Figure 8.

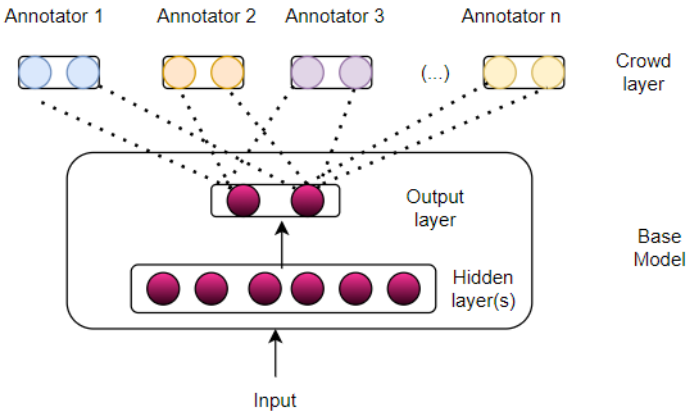


Figure 8: Crowd layer for the image classification task (Rodrigues and Pereira, 2018)

To experiment with the deep learning from crowds method, we added a crowd layer to the base models for each task (see Section 5.1) using the DL-MW variant that achieved the most accurate predictions in Rodrigues and Pereira (2018). In this variant, $f_r(\boldsymbol{\sigma})$ is defined as $\boldsymbol{W}^r(\boldsymbol{\sigma})$, where $\boldsymbol{W}^r$ is an annotator-specific matrix of the estimated sensitivities and specificities of the annotators, which we initialized to an identity matrix that is a trainable parameter of the neural network model. Like Rodrigues and Pereira (2018), we removed the crowd layer at test time and evaluated the model based on its softmax output.

5.2.4 Using both Hard Labels and Information about Disagreements

These methods use both the aggregated labels and additional information about disagreements. In our experiments, we used the gold label as an aggregate label, as proposed by both Plank et al. (2014a) and Fornaciari et al. (2021).

1. **Plank-style weighting** Plank et al. (2014a) proposed learning from annotator confusion by weighting the loss of each item (task example) by the inverse of how "confusing" the annotators found it. They characterized the confusion using two metrics: F1 *scores* between annotators on individual POS tags and *tag confusion probabilities* derived from confusion matrices they computed using the annotation of two expert annotators on 500 Twitter posts, distinct from the dataset to be trained using PEWE. For our experiment, we used the tag confusion probability, which was shown by Plank et al. to perform better than the F1-scores metric.

   To compute the tag confusion probabilities, we first generated a confusion matrix cm over all the POS tags. From this matrix, the probability of confusing two tags, t_1 and t_2, for a given item i was computed as the mean of the probability that annotator A_1 assigned one tag and A_2 another, and vice versa, i.e., $\{t_1, t_2\}$ is the mean of $cm[t_1, t_2]$ and $cm[t_2, t_1]$ (Plank et al., 2014a). Having computed these values for every pair of tags (labels), we augmented the loss function of the classifier by multiplying the loss for each item by $1 - \{y_g, y_p\}$, where y_g is the gold label for the given item and y_p is the predicted label (Plank et al., 2014a).

   We adapted this idea to a multi-annotator scenario using the multiple annotations collected for each dataset. First, for each item, we computed the confusion matrix between all pairs of annotators and calculated the average confusion matrix across all annotators. We then computed the average confusion matrix across all the items. We did this for each task independently. Using this matrix, we augmented the loss function of the base classifier for each task, following Plank et al. (2014a).

2. **Multi-task learning** A different approach to combining gold information with disagreement information was proposed by Fornaciari et al. (2021). Their model is based on multi-task learning (Caruana, 1997). Rather than having a single output layer, the model has two output layers; one for learning from an aggregated label and the other for learning from probabilistic soft labels. In this model, (henceforth, MTLSL), the total loss is the sum of the loss from comparing output 1 with the gold label and output 2 with the probabilistic label distribution. The first loss is computed using a negative log-likelihood loss function while the second (auxiliary) loss was learned using the KL divergence loss function. This method was found to improve the performance over using the gold label alone in several tasks including part-of-speech tagging.

   We also tested a second architecture also based on multi-task learning (henceforth, MTLOA). In this architecture the main task is again to learn the gold label, but the auxiliary task is to learn the item observed agreement. The loss for the auxiliary function is computed by calculating the mean-squared error between the predicted agreement (a Sigmoid squashed output of a neural network) and the observed agreement (Artstein and Poesio, 2008) for all items.

6. Results

In order to go beyond a simple listing of current approaches to training from disagreement as found in Section 5.2, we used each approach to train all the models discussed in Section 5.1 over each of the six datasets discussed in Section 2.6, evaluating the resulting models

using the metrics discussed in Section 4. The results are summarized in this section and analyzed in greater detail in the next two.

Tables 7 to 13 present the results for all the methods on all the tasks using a distinct evaluation metric. For comparison, we also include the results obtained by training the same architectures using gold labels (with a cross-entropy loss function). To account for non-deterministic model training effects, each model was re-run 30 times except for (i) IS, which was only run 10 times owing to the size of the dataset and model complexity, and (ii) IC-CIFAR10H and MRE, which were also run only 10 times due to model complexity. We report the average of these runs.

In these tables, there is one row for each method for learning from disagreement and one column for each dataset used in the evaluation. Double lines are used to group closely related methods in sections, one for each category of methods in Section 5.2 (e.g., aggregation methods). The best result for each dataset is highlighted in bold, and the best result among the training methods that do not use gold information is underlined. In each cell (i.e., for the result of a given training method on a particular dataset), we also include in superscript the row number of the method with the least significant improvement over the method in the cell, if any (significance is conducted via bootstrap sampling, following Berg-Kirkpatrick et al. (2012) and Søgaard et al. (2014)).

As discussed in Section 5.2, one category of methods we tested involves training a classifier using a soft loss function, i.e., using a standard loss function like cross-entropy, but targeting a probabilistic label generated from the crowd annotations instead of a one-hot label as is done, e.g., in Peterson et al. (2019) and Uma et al. (2020). There are, however, a number of ways in which such a probabilistic label could be obtained. Peterson et al. generated these probabilistic labels using a standard normalization function over the distribution of annotations for a given item. Uma et al. compared probabilistic soft labels estimated using standard normalization with probabilistic soft labels estimated using the softmax function. In this work, we further expanded this comparison by considering another natural way of generating such probabilistic labels, namely, using the posterior probability distribution obtained by two popular aggregations, D&S and MACE. The results of this analysis are discussed in detail in Section 7.1. The results in Tables 7 to 13 are those obtained using, for each method, the way of obtaining these labels that gave the best results.

6.1 Evaluation Against Gold or Hard Labels

Tables 7, 8, and 9 show the results of evaluation against gold labels using accuracy, F1, and the weighted version of F1 developed in the CrowdTruth project, CT F1. Figure 9 summarizes these results by displaying for each category of methods the results obtained by the best-performing approach in that category for each dataset.

The first broad conclusion we can reach from these tables and from Figure 9 is that the answer to Research Question **RQ3a** (Can methods for learning from disagreement that do not assume the existence of a gold label outperform methods that do?) is in most cases negative if we use "hard" evaluation: For three of the five datasets for which gold information is available for training (POS, IC-LABELME, and MRE), training using gold labels (alone, or in conjunction with crowd information) gave better results for hard evaluation

Table 7: Accuracy results for all methods on all tasks

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	89.08 [17]	NA	84.88	61.37	**97.18**	65.57 [7]
2	MV Silver	78.09 [9]	90.71 [5]	75.17 [12]	60.67 [1]	80.23 [4]	65.31 [7]
3	D&S Silver	77.67 [4]	92.80	75.20 [12]	60.37 [7]	83.58 [9]	65.65 [7]
4	MACE Silver	78.08 [9]	92.90	75.15 [12]	60.55 [1]	82.53 [6]	65.52 [7]
5	CrowdTruth Silver	79.33 [7]	91.30 [6]	75.17 [12]	60.37 [13]	84.50 [13]	64.09 [4]
6	SREL	79.23 [7]	92.11 [15]	75.66 [12]	60.01 [2]	83.46 [9]	**68.46**
7	CE loss + probabilistic labels	79.80 [1]	92.86	75.55 [12]	60.87	84.66 [13]	66.54 [10]
8	KL loss + probabilistic labels	79.96 [1]	92.86	75.53 [12]	60.68 [1]	84.73 [13]	66.58 [10]
9	MSE loss + probabilistic labels	79.20 [7]	92.90	75.50 [12]	60.70 [16]	84.21 [7]	63.49 [4]
10	DLC	77.87 [4]	92.82	74.67 [4]	59.75 [5]	83.69 [9]	**68.25**
11	MV + OA Hard Filter	72.20 [3]	68.51 [12]	74.85 [4]	54.77 [12]	86.05 [12]	63.98 [4]
12	Gold + OA Hard Filter	79.84 [1]	73.28 [14]	83.18 [1]	55.77 [10]	94.60 [16]	63.54 [4]
13	MV + OA Weighting	78.17 [9]	90.44 [6]	75.29 [12]	61.04	85.54 [12]	65.99 [10]
14	MV + WH Weighting	NA	90.31 [5]	75.25 [12]	58.76 [6]	NA	NA
15	Gold + PEWE	89.26 [17]	92.70	**85.43**	61.15	96.37 [17]	64.78 [13]
16	MTLOA	89.26 [17]	92.86	85.41	61.00	96.13 [17]	65.23 [7]
17	MTLSL	**90.11**	**92.95**	85.42	**61.43**	96.82 [1]	62.33 [5]

than training with crowd information alone, irrespective of which measure was used. In fact, the difference between the best method using gold and the best method only using crowd annotations could be quite large for these three datasets, up to 10 points in some cases (e.g., POS).

However, the answer to **RQ3a** is not entirely negative, because with RTE and IC-CIFAR10H it was the other way around: With IC-CIFAR10H, the best results were obtained using crowd information alone, and with RTE there was no significant difference between training with gold and training using silver labels aggregated with MV. Also, we anticipate that the situation will be completely reversed when soft evaluation metrics are employed; with this type of evaluation, using crowd information always improves results over only using gold labels, as shown in Section 6.2.

Another finding clearly emerging from the tables and the figure is that there is no evidence that the approach to using disagreement information that may appear most intuitive, filtering (i.e., using disagreement information to remove hard items) helps with hard evaluation. For none of these datasets were the best results obtained by filtering difficult items; on the contrary, filtering typically led to worse results, sometimes substantially so. The one exception is IC-LABELME: In this case the results obtained by filtering, while much worse than those obtained by using gold labels without filtering, were on par with those obtained with other ways of using crowd information.

A third observation is that the answer to **RQ2a** (What is the evidence that using crowd information helps in comparison to using only gold labels?) is mixed when using hard evaluation: Leveraging crowd information in addition to gold sometimes helps, although not by much, but other times it does not. With two of the five datasets for which we have

Table 8: F1 on all tasks using all methods

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	88.99 [17]	NA	84.46	**61.28**	**97.18**	65.54 [7]
2	MV Silver	76.86 [5]	90.55 [5]	65.24 [9]	60.63 [1]	79.52 [4]	65.13 [7]
3	D&S Silver	76.64 [2]	92.78	67.80 [5]	60.32 [15]	83.03 [9]	65.53 [7]
4	MACE Silver	77.08 [5]	92.87	65.28 [9]	60.45 [15]	81.87 [6]	65.40 [7]
5	CrowdTruth Silver	78.14 [7]	91.13 [6]	76.11 [12]	59.52 [3]	83.99 [13]	63.90 [4]
6	SREL	78.21 [7]	92.00 [15]	67.19 [5]	58.66 [5]	82.96 [9]	**68.36**
7	CE loss + probabilistic labels	78.75 [1]	92.84	66.44 [3]	60.68	84.02 [13]	66.43 [10]
8	KL loss + probabilistic labels	78.92 [1]	92.84	66.44 [3]	60.43 [15]	84.09 [13]	66.45 [10]
9	MSE loss + probabilistic labels	78.14 [7]	92.88	66.38 [3]	60.51 [15]	83.61 [13]	63.33 [4]
10	DLC	76.27 [2]	92.74	63.87 [2]	58.42 [5]	83.19 [9]	**67.99**
11	MV + OA Hard Filter	68.85 [10]	57.56	64.34 [2]	46.76 [12]	85.37 [12]	63.69 [4]
12	Gold + OA Hard Filter	76.99 [9]	65.50	82.38 [1]	49.55 [10]	94.59 [16]	63.17 [15]
13	MV + OA Weighting	76.86 [9]	90.21 [5]	65.16 [9]	60.74	84.88 [12]	65.89 [10]
14	MV + WH Weighting	NA	90.13 [5]	65.34 [9]	58.53 [6]	NA	NA
15	Gold + PEWE	89.18 [17]	92.65	**85.07**	61.12	96.37 [17]	64.67 [13]
16	MTLOA	89.15 [17]	92.82	84.95	60.99	96.13 [17]	65.18 [7]
17	MTLSL	**90.06**	**92.92**	84.87	61.13	96.82 [1]	62.34 [15]

Table 9: CrowdTruth weighted F1 for all tasks using all methods

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	92.46 [17]	NA	86.94	**74.05**	**98.25**	78.48 [10]
2	MV Silver	85.40 [6]	94.54 [5]	70.02 [7]	73.39 [15]	87.33 [4]	78.14 [7]
3	D&S Silver	85.27 [2]	96.00	75.34 [5]	73.24 [15]	89.23 [9]	78.50 [10]
4	MACE Silver	85.69 [6]	96.02	70.13 [7]	73.46 [1]	88.80 [6]	78.43 [10]
5	CrowdTruth Silver	86.58 [7]	94.84 [6]	82.17 [12]	72.15 [4]	90.11 [7]	77.21 [4]
6	SREL	86.51 [7]	95.43 [10]	74.89 [5]	71.19 [5]	89.48 [9]	**80.56**
7	CE loss + probabilistic labels	87.15 [1]	96.03	72.80 [6]	73.40 [1]	90.17 [13]	79.17 [10]
8	KL loss + probabilistic labels	87.27 [1]	96.01	73.10 [6]	73.17 [16]	90.20 [13]	79.09 [10]
9	MSE loss + probabilistic labels	86.61 [7]	96.04	73.21 [6]	73.27 [15]	89.90 [13]	76.74 [15]
10	DLC	84.76 [2]	95.87 [3]	66.11 [2]	71.06 [5]	89.57 [9]	**80.30**
11	MV + OA Hard Filter	78.54 [12]	67.76 [12]	66.47 [2]	59.20 [12]	91.04 [12]	77.11 [4]
12	Gold + OA Hard filter	82.96 [3]	74.18 [6]	84.76 [1]	62.12 [10]	96.47 [16]	76.65 [15]
13	MV + OA Weighting	85.31 [6]	94.26 [2]	70.63 [7]	73.37 [15]	90.76 [12]	78.80 [10]
14	MV + WH Weighting	NA	94.29 [2]	70.26 [7]	71.63 [8]	NA	NA
15	Gold + PEWE	92.60 [17]	95.87	87.53	73.87	97.76 [17]	77.76 [13]
16	MTLOA	92.56 [17]	95.98	87.42	73.83	97.60 [17]	78.17 [7]
17	MTLSL	**93.07**	**96.05**	**87.65**	73.66	98.02 [1]	76.00 [15]

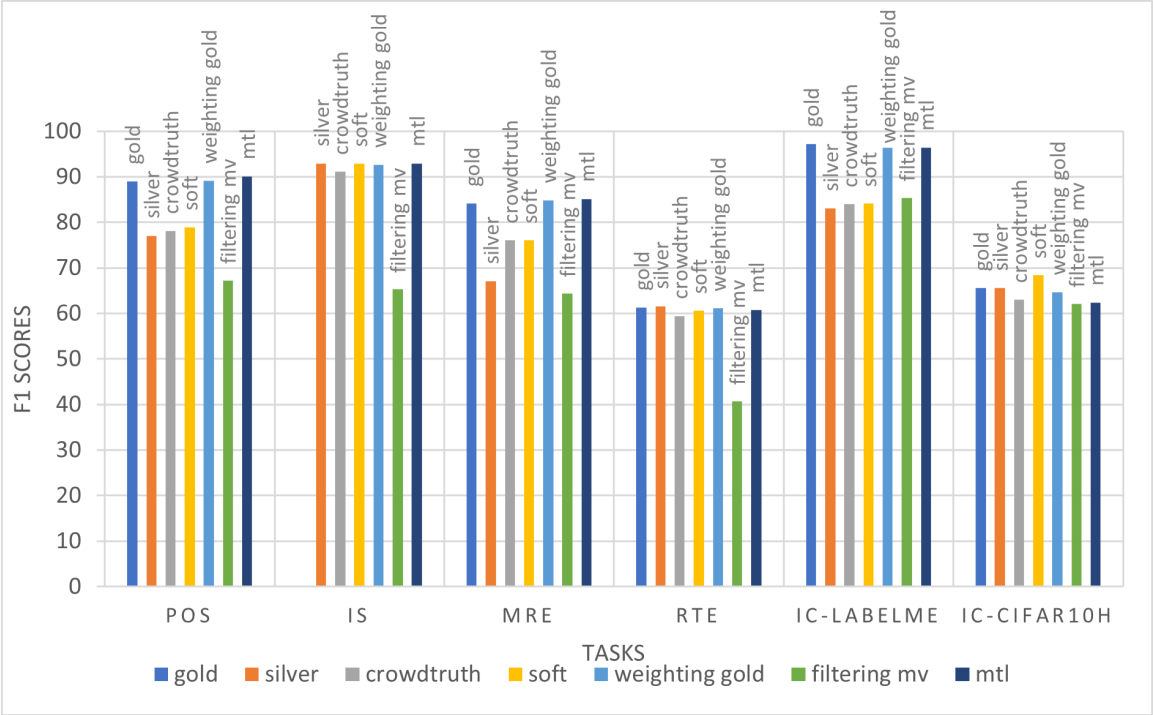


Figure 9: Graph showing the F1 scores of the best-performing training approach for each category on all the datasets

gold information—POS and MRE—using information from the soft label to supplement the gold according to the MTLSL method improved performance over using only gold labels with all three hard metrics. This difference was small—typically around one percentage point—but significant in the case of POS. With RTE, the best-performing method depended on the metric, but the differences were never significant.[26] With the two IC datasets, however, using gold labels alone yielded a significant improvement in the results only when compared to using gold labels in combination with crowd information, especially in the case of the IC-CIFAR10H dataset. (Although we should quickly qualify that statement by adding that that doesn't mean crowd information is not useful with IC-CIFAR10H—on the contrary, the best overall results were obtained using crowd information only!) Using crowd information in addition to the hard label also helped slightly in the IS task when the hard label was an aggregated silver label, although the difference was not significant. Again, we must immediately point out that the situation was reversed with soft evaluation.

The answer to **RQ2b** (What is the best way to leverage crowd information in addition to gold labels?) is that with most datasets, MTLSL is either significantly better, better, or indistinguishable from other approaches. The one exception is IC-CIFAR10H, where MTLSL performed rather poorly—but in this case the best among the approaches leveraging both gold labels and crowd information was the other multi-task learning approach we tested, MTLOA, which uses observed agreement as an auxiliary function.

26. In line with Card et al. (2020), who discuss statistical power in relation to dataset size, it might be worth considering retiring this small dataset.

The final observation is that the answer to **RQ3b** (Which method for learning from disagreement achieves the best results?) is that there isn't a clear "winner" among the methods not using a gold label: What methods achieve the best results depends on the dataset.

We summarize the results as follows, using $A >> B$ to signify that most methods in category A are significantly better than most methods in category B, $A \sim B$ to signify that most methods in category A are statistically indistinguishable from most methods in category B, and $A \geq B$ to signify that some methods in category A are significantly better than some methods in category B, whereas others are equivalent.

1. On POS, the best results among the methods using only crowd information were obtained by the three methods using a soft loss function, then by using aggregation, then weighting and filtering. The performance ranking for POS shown in Figure 9 can be schematically summarized as follows, where $HARD_{GOLD}$ is gold training, $SOFT$ includes the CrowdTruth method, and categories are ranked by the performance of the best-performing method in the category:
$AUGMENTED_{GOLD} >> HARD_{GOLD} >> SOFT \geq FILTER_{GOLD} >>$
$HARD_{SILVER} \geq WEIGHT_{SILVER} >> FILTER_{SILVER}$

2. On PDIS, no gold labels are available, so the silver label achieving the best results (aggregated with MACE) was used as the hard label. The best results for this hard silver label were obtained using MTLSL, but augmenting hard silver with crowd information, using hard silver only, or using crowd information only with soft-label methods achieved statistically indistinguishable results on this dataset. The only significant differences were between any of these methods and weighting and between weighting and hard filtering, which gave extremely poor results.
$AUGMENTED_{SILVER} \sim HARD_{SILVER} \sim SOFT >> WEIGHT_{SILVER} >>$
$FILTER$

3. MRE is the one dataset on which different methods achieved the best results depending on which hard evaluation metric was used, for reasons discussed in more detail in the following section. Methods exploiting both gold labels and crowd information achieved the best results with all three hard metrics, systematically outperforming training with gold only, although the difference was not significant. But among the methods not relying on gold labels, CrowdTruth aggregation obtained by far the best results in terms of F1 and especially of CT F1, with a margin of 10 points or more over other methods. Soft-label methods achieved the best accuracy results, although the difference was not significant.
$AUGMENTED_{GOLD} \sim HARD_{GOLD} >> FILTER_{GOLD} >> CT \sim_{ACCURACY}$
$/ >>_{F1} SOFT \sim_{ACCURACY} / >>_{F1} HARD_{SILVER} \geq WEIGHT_{SILVER}$
$\geq FILTER_{SILVER}$

4. RTE is one of the two datasets for which using gold labels did not yield better results than using crowd information only. The results using gold labels, gold labels augmented with crowd information, silver weighting, and some of the soft loss functions were all statistically equivalent. Among the soft-labelling methods, the best results

were obtained by OA weighting, then soft loss using CE, then aggregation. But all methods achieved roughly comparable results with all metrics, with a maximum 1-2 percentage points between the worse and the best results; again, the only exception is hard filtering, which performed substantially worse.
$AUGMENTED_{GOLD} \sim HARD_{GOLD} \sim WEIGHT_{SILVER} \sim SOFT >>$
$HARD_{SILVER} \sim>> FILTER$

5. The best results with IC-LABELME were obtained using gold labels alone (which did very slightly, but significantly, better than combining gold labels with crowd information). The next best results were obtained using OA for filtering or weighting silver labels—this is the only dataset in which filtering/weighting silver items proved to be a competitive approach. Soft labels were next, then aggregation. Using hard silver labels yielded the worst results in terms of hard evaluation metrics, but this is the dataset on which probabilistic aggregation outperformed MV by the largest margin: training over the CT-aggregated labels, while not resulting in the best F1, improved performance over training with the MV labels by more than 4 points.
$HARD_{GOLD} >> AUGMENTED_{GOLD} >> FILTER_{GOLD} >> FILTER_{SILVER} \sim$
$WEIGH_{SILVER} >> SOFT >> HARD_{SILVER}$

6. Finally, IC-CIFAR10H was the one dataset for which using crowd information yielded significantly better results only when compared to using gold, or augmented gold, labels. The best results were obtained using SREL—an improvement of around three points—but soft-loss training also significantly outperformed gold training, which was statistically indistinguishable from both silver training and MTLOA.
$SOFT >> WEIGH_{SILVER} >> HARD_{SILVER} \sim HARD_{GOLD} >>$
$AUGMENTED_{GOLD} >> FILTER$

We further analyze these results on a task-by-task basis in Section 7, aiming to explain these dataset-dependent differences.

One final consideration: It can be observed that the three evaluation metrics tend to be aligned, in the sense that the methods performing best on a given task were the same irrespective of the evaluation used, with the few exceptions noted.

6.2 Evaluation Against Soft Labels

Given the empirical evidence challenging the assumption that it is always possible to assign a unique label to items in cognitive tasks reviewed in Section 2, the form of evaluation discussed in the previous subsection—testing models against gold labels—while standard in NLP and in AI, is at the very least questionable. In this paper, therefore, we also used the "soft" evaluation metrics discussed in Section 4 to analyze current methods for training with disagreement. The results are shown in Tables 10 to 13.

Arguably, the main result of this paper is that the answer to **RQ3a**, which as seen in Section 6.1 is mainly negative when using hard evaluation metrics, becomes positive with soft evaluation metrics, i.e., the ranking among methods for learning from disagreement seen in the previous section is to a large extent reversed when these methods are evaluated using a soft evaluation metric, so that methods *not* using gold labels generally outperform hard-training methods for all datasets and all metrics.

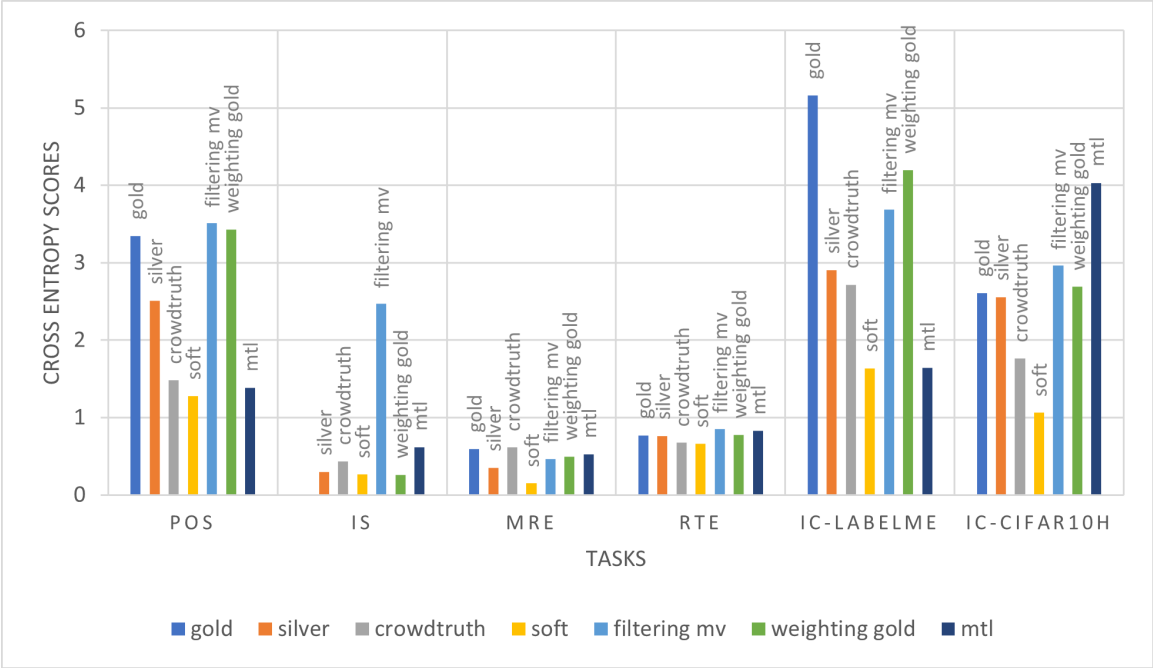


Figure 10: Graph showing the cross entropy scores of the best performing training approach for each category on all the datasets (lower is better)

The answer to **RQ3b**—which of these methods performs best—again depends on the task and, to a lesser extent, on the metric, but for almost all metrics and almost all tasks the best results were obtained by some form of soft-loss training or repeated labelling.

The answer to **RQ2a** is also uniformly positive: With soft evaluation, using crowd information always helps improve the results over training using gold labels only, for all evaluation metrics and all datasets. In answer to **RQ2b**, some form of multi-task learning with an auxiliary function capturing disagreements is usually the best approach, with MTLSL in particular achieving fair results in many cases.

On a task-by-task basis, the results can be summarized as follows (a more detailed discussion can be found in the next section):

- Almost all training methods using gold labels, except for MTLSL, achieved significantly worse performance under all soft metrics with POS, as with all other datasets. Soft-loss training methods performed best, with CE- and KL-loss training performing significantly better than all other methods according to all soft evaluation metrics. Interestingly, MTLSL performed better than most soft-labelling methods other than soft loss, but MTLOA did not. It is also worth noting that CrowdTruth aggregation achieved better results than the other aggregation methods. This can be loosely summarized as follows:
  $SOFT_{CE,KL} \sim_{CE} / >>_{JSD,CS} MTLSL \geq CT \sim SOFT \geq HARD_{SILVER} >> WEIGHT \sim FILTER >> HARD_{GOLD}$

Table 10: Cross entropy between produced probabilities and soft labels for all tasks using all methods (smaller is better)

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	3.346 [16]	NA	0.574 [17]	0.771 [8]	5.159 [12]	2.607 [5]
2	MV Silver	2.583 [4]	0.397 [14]	0.522 [11]	0.785 [3]	3.065 [4]	2.627 [5]
3	D&S Silver	2.524 [4]	0.300 [9]	0.350 [6]	0.772 [8]	2.902 [10]	2.554 [5]
4	MACE Silver	2.506 [10]	0.297 [9]	0.460 [3]	0.797 [13]	2.906 [10]	2.646 [5]
5	CrowdTruth Silver	1.482 [9]	0.403 [14]	0.610 [16]	0.673 [6]	2.717 [6]	1.763 [9]
6	SREL	1.787 [5]	0.359 [9]	0.310	0.669	2.572 [9]	**1.062**
7	CE loss + probabilistic labels	**1.358**	0.273 [16]	0.310	0.740 [9]	**1.638**	**1.112**
8	KL loss + probabilistic labels	**1.279**	0.265 [16]	0.309	0.742 [9]	**1.638**	**1.109**
9	MSE loss + probabilistic labels	1.442 [7]	0.289 [8]	**0.309**	0.717 [5]	1.747 [7]	1.491 [8]
10	DLC	2.136 [6]	0.275 [16]	0.715 [15]	**0.668**	2.798 [5]	3.507 [11]
11	MV + OA Hard Filter	3.243 [15]	2.246 [12]	0.490 [4]	0.879 [14]	3.684 [13]	2.961 [15]
12	Gold + OA Hard Filter	3.115 [13]	1.863 [17]	0.495 [4]	0.879 [14]	4.612 [15]	2.844 [15]
13	MV + OA Weighting	2.759 [2]	0.372 [6]	0.527 [14]	0.787 [3]	3.121 [2]	2.615 [5]
14	MV + WH Weighting	NA	0.379 [6]	0.516 [11]	0.842 [4]	NA	NA
15	Gold + PEWE	3.432 [1]	0.261 [16]	0.621 [16]	0.779 [3]	4.198 [16]	2.691 [15]
16	MTLOA	3.288 [12]	**0.245**	0.579 [17]	0.796 [13]	3.926 [11]	2.505 [5]
17	MTLSL	**1.382**	0.618 [5]	0.569 [13]	0.786 [3]	**1.642**	4.032 [1]

Table 11: Jensen-Shannon divergence results on all tasks using all methods (smaller is better)

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	0.413 [11]	NA	0.251 [12]	0.415 [7]	0.547 [12]	0.405 [10]
2	MV Silver	0.353 [4]	0.218 [16]	0.166 [4]	0.416 [7]	0.452 [4]	0.399 [10]
3	D&S Silver	0.353 [4]	0.129 [17]	0.156 [7]	0.416 [7]	0.449 [5]	0.404 [10]
4	MACE Silver	0.351 [10]	0.129 [17]	0.163 [3]	0.415 [7]	0.448 [10]	0.395 [10]
5	CrowdTruth Silver	0.236 [7]	0.243 [13]	0.297 [15]	0.425 [12]	0.428 [6]	0.417 [1]
6	SREL	0.318 [9]	0.268 [10]	**0.136**	0.426 [11]	0.417 [9]	0.415 [1]
7	CE loss + probabilistic labels	**0.207**	0.146 [9]	0.148 [6]	0.413 [13]	**0.201**	0.427 [1]
8	KL loss + probabilistic labels	**0.206**	0.150 [7]	0.148 [6]	0.413 [13]	**0.201**	0.428 [1]
9	MSE loss + probabilistic labels	0.280 [17]	0.128 [17]	0.148 [6]	0.416 [7]	0.208 [7]	0.431 [1]
10	DLC	0.342 [6]	0.220 [2]	0.177 [17]	0.426 [11]	0.430 [6]	**0.368**
11	MV + OA Hard Filter	0.397 [13]	0.351 [12]	0.169 [14]	0.430 [12]	0.489 [13]	0.407 [2]
12	Gold + OA Hard Filter	0.426 [15]	0.303 [5]	0.241 [10]	0.421 [1]	0.539 [15]	0.412 [2]
13	MV + OA Weighting	0.353 [4]	0.232 [10]	0.166 [4]	0.410 [17]	0.464 [2]	0.392 [10]
14	MV + WH Weighting	NA	0.256 [10]	0.167 [4]	0.422 [16]	NA	NA
15	Gold + PEWE	0.415 [11]	0.163 [8]	0.258 [1]	0.417 [1]	0.537 [16]	0.407 [2]
16	MTLOA	0.413 [11]	0.178 [15]	0.250 [12]	0.418 [1]	0.531 [11]	0.401 [10]
17	MTLSL	0.236 [7]	**0.096**	0.172 [11]	**0.404**	**0.201**	0.415 [1]

Table 12: Cosine Similarity between the entropy of the produced distribution and the annotation label distribution for all tasks using all methods

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	0.659 [11]	NA	0.655 [12]	0.567 [3]	0.551 [16]	0.389 [5]
2	MV Silver	0.758 [10]	0.115 [13]	0.478 [4]	0.570 [7]	0.778 [3]	0.383 [5]
3	D&S Silver	0.762 [10]	0.176 [9]	0.700 [7]	0.571 [7]	0.797 [10]	0.391 [5]
4	MACE Silver	0.750 [10]	0.183 [9]	0.548 [15]	0.560 [2]	0.777 [3]	0.379 [5]
5	CrowdTruth Silver	0.885 [7]	0.116 [13]	0.717 [7]	0.589 [6]	0.840 [10]	0.472 [9]
6	SREL	0.873 [7]	0.167 [4]	**0.772**	**0.590**	0.860 [17]	0.546
7	CE loss + probabilistic labels	**0.899**	0.204 [15]	0.761 [6]	0.579 [9]	**0.979**	0.546
8	KL loss + probabilistic labels	**0.907**	0.211 [15]	0.763 [6]	0.579 [9]	0.978 [7]	**0.547**
9	MSE loss + probabilistic labels	0.888 [7]	0.191 [7]	0.761 [6]	0.584 [5]	0.978 [7]	0.506 [6]
10	DLC	0.849 [6]	0.207 [15]	0.688 [3]	0.589 [6]	0.852 [6]	0.331 [13]
11	MV + OA Hard Filter	0.698 [13]	0.065 [5]	0.590 [15]	0.517 [4]	0.697 [14]	0.390 [5]
12	Gold + OA Hard Filter	0.729 [4]	0.071 [5]	0.678 [5]	0.518 [4]	0.592 [12]	0.379 [5]
13	MV + OA Weighting	0.720 [4]	0.136 [14]	0.455 [2]	0.570 [7]	0.755 [4]	0.372 [5]
14	MV + WH Weighting	NA	0.161 [4]	0.501 [4]	0.571 [7]	NA	NA
15	Gold + PEWE	0.650 [16]	0.241 [16]	0.641 [1]	0.570 [7]	0.597 [16]	0.391 [5]
16	MTLOA	0.667 [11]	**0.264**	0.655 [12]	0.567 [3]	0.625 [11]	0.393 [5]
17	MTLSL	0.876 [5]	0.071 [5]	0.433 [2]	0.563 [1]	0.976 [7]	0.352 [13]

Table 13: Pearson correlation between the entropy of the produced distribution and the annotation label distribution for all tasks using all methods

		POS	PDIS	MRE	RTE	IC-LABELME	IC-CIFAR10H
1	Gold	0.399 [11]	NA	0.223 [4]	0.037 [7]	-0.016 [2]	0.127 [5]
2	MV Silver	0.517 [10]	-0.104 [11]	0.214 [4]	0.043 [7]	0.026 [11]	0.118 [5]
3	D&S Silver	0.504 [10]	0.029 [9]	0.382 [7]	0.039 [7]	0.111 [11]	0.125 [5]
4	MACE Silver	0.513 [10]	0.032 [9]	0.265 [3]	0.022 [1]	0.139 [11]	0.112 [5]
5	Crowd Truth Silver	0.642 [7]	-0.113 [2]	0.293 [7]	0.037 [7]	0.194 [10]	0.160 [9]
6	SREL	0.635 [7]	-0.098 [10]	**0.511**	0.030 [2]	0.284 [8]	0.217
7	CE loss + probabilistic labels	**0.656**	0.051 [15]	0.444 [6]	0.056	0.407 [9]	0.215
8	KL loss + probabilistic labels	**0.663**	0.053 [15]	0.450 [6]	0.059	0.403 [9]	**0.217**
9	MSE loss + probabilistic labels	0.640 [7]	0.047 [15]	0.435 [6]	0.058	**0.425**	0.190 [6]
10	DLC	0.603 [6]	-0.040 [17]	0.010 [17]	0.025 [3]	0.263 [6]	0.119 [5]
11	MV + OA Hard Filter	0.411 [12]	-0.023 [17]	0.208 [4]	-0.061 [12]	0.192 [10]	0.121 [5]
12	Gold + OA Hard Filter	0.451 [6]	-0.029 [17]	0.227 [3]	-0.046 [4]	0.130 [11]	0.107 [5]
13	MV + OA Weighting	0.517 [10]	-0.102 [10]	0.211 [4]	0.056	0.195 [10]	0.100 [5]
14	MV + WH Weighting	NA	-0.091 [10]	0.217 [4]	**0.065**	NA	NA
15	Gold + PEWE	0.395 [11]	0.067	0.217 [4]	0.030 [3]	-0.020 [2]	0.129 [5]
16	MTLOA	0.408 [12]	**0.079**	0.215 [4]	0.035 [7]	-0.016 [2]	0.124 [5]
17	MTLSL	0.612 [6]	0.020 [7]	0.120 [13]	0.047 [7]	0.376 [7]	0.105 [5]

- With PDIS, MTL methods (using silver as the hard label) performed best according to all four soft evaluation metrics, but the type of MTL that worked best depended on the evaluation, and sometimes the difference in results was quite substantial. E.g., with cross entropy, MTLOA was the best type of training, but MTLSL was the worst. Soft-loss methods were next best, then methods that rely on a prior aggregation. It should be noted that the best results with soft-loss functions with this dataset were obtained using the posterior of probabilistic aggregation methods as the target, suggesting that filtering noise from crowd annotations helps with this dataset. It should also be noted that all methods struggled to predict the entropy of the annotator label distribution with this dataset, whether computed using cosine similarity or, even more so, using Pearson correlation, again suggesting that there is lots of noise.
$MTLOA/MTLSL >> SOFT >> HARD_{SILVER} \geq WEIGHT \geq FILTER$.

- With MRE, soft-labelling methods performed best, but different types of training achieved the best results depending on the evaluation used. Repeated labelling generally performed best, followed by soft-loss methods, except for cross-entropy, where it was the other way around; however, the difference was typically not significant.
$SOFT_{SHENG} >> SOFT_{LOSS} \geq HARD_{SILVER} \geq FILTER, WEIGHT \geq AUGMENTED_{GOLD} \geq HARD_{GOLD}$.

- One striking aspect of the results with RTE is that those for the different methods were much closer than with other datasets, although significant differences did emerge. In particular, while the methods relying only on crowd information outperformed gold training and training on aggregated silver labels according to most soft metrics, the differences were much smaller, and MTLSL outperformed the soft-labelling methods in terms of JSD. For this dataset, SREL and DLC achieved the best results in terms of cross-entropy (there was a small difference between the two, but it was not significant) and cosine similarity; the difference with other soft-loss methods was significant. Another noticeable result is that the Pearson correlation between the entropy of the produced distribution and that of the target distribution was mostly near 0. The results with this dataset are difficult to summarize because the soft metrics did not all point to the same ranking, but as a first approximation, we can say that:
$SOFT_{SHENG,DLC} \geq CT \geq MTLSL \geq HARD_{SILVER} \geq WEIGHT, FILTER \geq HARD_{GOLD}$.

- With IC-LABELME, the best-performing method on all metrics was training using a soft loss function with softmax distribution, although again MTLSL was, in most cases, a very close second and was equivalent to soft-loss training when evaluated using cross entropy or JSD. Unsurprisingly perhaps, for this task, using only gold labels for training resulted in a extremely poor match regarding predicted entropy.
$SOFT \geq MTLSL >> CT \geq HARD_{SILVER} >> WEIGHT >> FILTER \geq HARD_{GOLD}$.

- And finally, for IC-CIFAR10H, soft-labelling methods clearly performed better than hard, although, again, which method performed best—soft-loss, repeated labelling—depended on the measure used.
$SOFT >> CT, HARD_{SILVER} \geq FILTER, WEIGHT \geq HARD_{GOLD}$.

6.3 Preliminary Discussion

In this section we saw that generally speaking, methods relying on hard labels (gold or silver) performed better when performance was measured using "hard" evaluation measures, whereas methods not assuming that such labels can be found performed best with "soft" evaluation. But the fact that the results were very much dataset dependent suggests that the performance of these methods is likely affected by the characteristics of the dataset. In the next section we carry out a detailed analysis exploring this suggestion.

7. A Dataset-by-Dataset Analysis

We just saw in Section 6 that the relative performance of current methods for learning from disagreement varies greatly from dataset to dataset with both hard and soft evaluation metrics. The aim of this section is to analyze these differences in greater depth, looking at each dataset in isolation in order to understand how the pattern of results observed in that dataset relate to its characteristics. Each subsection includes sections devoted to the results obtained on a dataset by training with gold labels supplemented with crowd information, with aggregated labels, and with soft labels only.

7.1 Generating Probabilistic Labels from Crowds

As anticipated in Section 6, the performance of soft-loss training methods depends very much on how the probabilistic labels are obtained. In that section, soft evaluation and soft-loss training was carried out using the best-performing probabilistic label. As a necessary preamble to the analysis that follows, we begin this section by discussing in detail how these "best labels" were determined.

Table 14: Different methods for generating probabilistic labels from crowd annotations and their effect on accuracy

	POS	IS	MRE	RTE	IC-LABELME	IC-CIFAR10H
Standard Norm	78.99 ± 0.36	90.68 ± 0.43	**75.79 ± 0.29**	60.24 ± 0.99	83.46 ± 0.82	**66.54 ± 0.95**
Softmax	**80.03 ± 0.28**	90.50 ± 0.55	75.27 ± 0.18	**60.87 ± 0.84**	**84.66 ± 0.52**	65.50 ± 1.10
D&S posterior	77.95 ± 0.61	92.74 ± 0.22	74.78 ± 0.26	60.51 ± 0.86	83.27 ± 0.76	65.16 ± 1.34
MACE posterior	78.27 ± 0.94	**92.81 ± 0.26**	75.32 ± 0.36	60.53 ± 0.83	83.53 ± 0.56	65.28 ± 1.02

The results in Table 14 illustrate the effect on accuracy of these different ways of obtaining the probabilistic labels. As we can see from that table, how the probabilistic distribution is obtained does affect the results. The results for POS, IC-LABELME, and IC-CIFAR10H are consistent with the findings in Uma et al. (2020): Estimating probabilistic labels using standard normalization is preferable for IC-CIFAR10H, while estimating the labels using softmax is preferable for POS and IC-LABELME. In addition, we found that using softmax gave the best results for the RTE dataset, standard normalization was best for MRE, and using the MACE posterior was best for PDIS. These differences, we hypothesize, can be attributed to the fact that the standard normalization function does not change the class proportions (as the softmax function does) or under-count disagreement (as the MACE and D&S posteriors

do) but retains the richness of the original representation. The differences between the datasets explain why these properties of these functions matter.

Like Uma et al. (2020), we hypothesize that the standard normalization function is a better choice for high-agreement datasets that also have a large distribution of good-quality annotations. This hypothesis is supported by the results from the tasks trained using datasets that meet this criteria: IC-CIFAR10H and MRE. For these datasets, which are characterized by a combination of (1) relatively higher observed agreement of 0.92 and 0.86, respectively; (2) a median of 50 and 15 annotators per item, respectively; (3) annotators with an average accuracy of 0.95 and 0.76, respectively; and (4) a majority of good-quality annotators (see Table 1), soft-loss training targeting standard normalization probabilistic labels yielded the most accurate results. In general, the trend seems to be that the higher the observed agreement, the higher the accuracy of training by targeting standard normalized soft labels over targeting softmaxed soft labels (see Figure 11).

In contrast, the softmax yields the best probabilistic label for low-agreement datasets, as it exacerbates disagreement and assigns a mass to to all items, even ones receiving no annotations. This affects performance with some datasets. Consider the following example from the POS dataset with the token to be tagged in bold:

> **Sentence**:"Journalists and Social Media experts alike will appreciate ***this*** spoof out of Dallas : URL"
> **Gold Label**: Determinant
> **Crowd annotations**: {Noun: 1, Pronoun:1, Adjective:1, Adposition:2}

The observed agreement for the item is 0.1, indicating that annotators found the item confusing. The standard normalization only assigns a probability to the four labels produced by annotators: {*Noun:0.2, Pronoun:0.2, Adjective:0.2, Adposition:0.4*}. Softmax, in addition to assigning probabilities to these four labels—{*Noun:0.11, Pronoun:0.11, Adjective:0.11, Adposition:0.31*}—also assigns a small probability of 0.04 to each of the other labels not selected by any annotators (including the *Determinant* class). So for this low-agreement item, although normalization and softmax produce distributions with the same mode (i.e., the majority vote), the softmax function (1) assigns a smaller mass to the modal class (which according to the gold standard is not the correct label for that item) and (2) assigns a small mass to the class chosen by the experts. Thus, for datasets like RTE that have a relatively low observed agreement of 0.63 and datasets like POS and IC-LABELME that, in addition to having relatively low agreement of 0.73, have no annotated gold labels for over 11% of items, the softmax function proves to be the best option.

The PDIS dataset is a mixed bag, with an observed agreement closer to MRE and IC-CIFAR10H than to that of POS and IC-LABELME; the results reflect this. The difference in accuracy between training with standard-normalized soft labels and training with softmaxed soft labels was smallest for the PDIS dataset. However, soft-loss methods for this task benefit from using the MACE and D&S models that try to discriminate between annotators and eliminate noise, likely because of the relatively lower observed agreement (leaving room for improvement) and high number of annotations per annotator (ample examples from which to learn annotator characteristics).

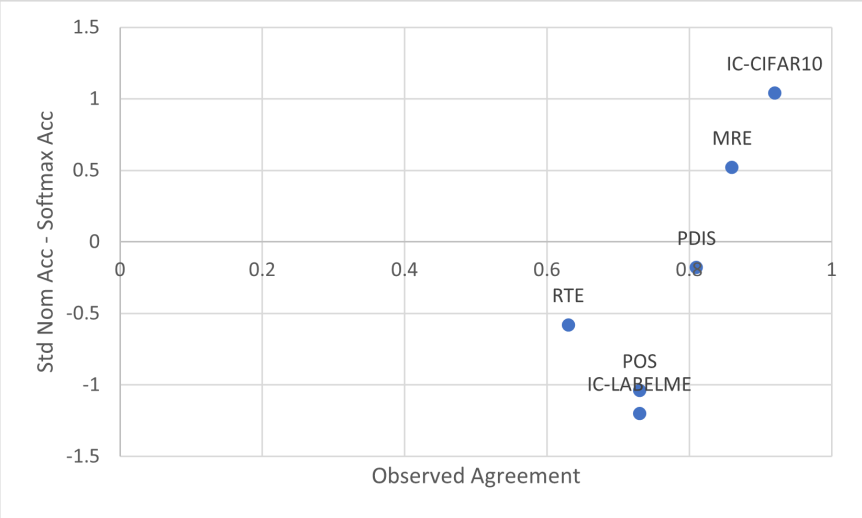


Figure 11: Graph of observed agreement against the difference in accuracy training with standard normalization (stdn) soft labels and training with softmax soft labels.

As a result of this analysis, in our experiments the soft labels used were standard-normalized soft labels for IC-CIFAR10H and MRE, MACE posterior for IS, and softmax soft labels for POS, RTE, and IC-LABELME.

7.2 Part-of-Speech Tagging

The key characteristics of the POS dataset (see Tables 1 and 2) are that it has the second highest number of items (14,000), average coder accuracy is high (.93), and the mean number of annotations per item is also fairly high (16.37). On the other end, observed agreement is relatively low (.73), and the quality of aggregated labels is low as well. Finally, while the raw annotator entropy is fairly low (.13), the best distribution entropy (BDE) is fairly high (.39).

7.2.1 Gold vs. Non-gold

As discussed in Section 6, with the POS dataset we saw clear differences in performance between the models using gold labels and those using silver or soft labels, in both directions. Methods using gold labels clearly did better in terms of hard evaluation metrics, although using soft information helped, whereas methods only using soft labels clearly did better in terms of soft evaluation metrics.

This substantial difference between using and not using gold labels is surprising given the high coder accuracy and the substantial number of annotations per item and the well-known findings of, e.g., Snow et al. (2008) and Sheng et al. (2008) that the quality of labels produced by the crowd is comparable to that of labels produced by experts provided that sufficient coders of sufficient quality are employed. To understand this surprising result, we carried out a more in-depth analysis of the data, the results of which are shown in Table 15. First of all, the table shows that the high mean number of annotators per item is deceptive: Whereas several items have a high number of annotations (177), for others the number of

annotations is much lower, so that the median of this figure is only 5. Second, Table 15 shows that this dataset is not uniform, but can be partitioned into two subsets with very different characteristics. Eighty percent of the judgments in the dataset are about nouns, even though, according to the gold labels, nouns constitute only 27.7% of the total number of items. The alignment between coders and gold labels (coder accuracy) for nouns is very high, almost 98%, and so is the average number of annotations per item. In contrast, only 20% of the judgments are about the remaining 72% of items, and "coder accuracy" on these is much lower. This suggests that, for the great majority of items, the agreement of the crowd with the gold labels is not high enough to ensure that training on soft labels only will achieve high accuracy when evaluated on gold. As discussed in Section 2.8.3, the analysis by Plank et al. (2014b) suggests that one reason for the low coder accuracy on this dataset is that many items are linguistically debatable, and the great majority of these cases involve categories other than nouns—although label overlap would also appear to play an important role and we cannot exclude some cases of difficulty or indeed annotator error (see Section 2.8). As we will see, Snow et al.'s hypothesis holds with datasets where coders agree much more with gold labels and where there are high number of annotations per item.

Table 15: Nouns vs non-nouns in the POS dataset

	Nouns	Others
Percentage of items in the subset	27.72	72.28
Percentage of judgments in category per	80.13	19.87
Average number of annotations per item	12.57	3.80
Average coder accuracy	97.89	69.08
Average item observed agreement	0.804	0.695
MV aggregated label accuracy	85.94	77.52

7.2.2 Using Soft Labels to Supplement Gold Labels

The second finding highlighted by Tables 7 to 9 is that using crowd information in addition to gold does help with this dataset. MTLSL stands out as the best method for learning hard or weighted truth for the POS dataset under all three hard evaluation metrics. This method, which targets the soft-label distribution as an auxiliary task to supplement the gold labels, achieved +1.03, +1.06, and +0.61 significant points over training on gold alone when evaluated using accuracy, F1, and CT F1 respectively. The other methods that augment gold labels with information from the crowd, MTLOA and PEWE, also outperformed the gold, although not by a significant margin.

The fact that the training methods leveraging crowd information improved over gold training suggest that the crowd provides information that usefully supplements the gold labels. As previously mentioned, the POS dataset is characterized by a combination of a relatively high number of judgments per item, accurate coders, relatively low observed agreement between them, and relatively high "best distribution entropy." It would seem then plausible that it is the quality, quantity, *and diversity* of crowd judgments that leads to the crowd information improving performance over gold labels—which provides further evidence that the low agreement is not so much due to poor "coder accuracy," but to the

fact that more than one interpretation is possible for several items in this dataset, as pointed out by Plank et al. (2014b). As we will see, this hypothesis that these are the conditions under which using soft labels in addition to gold labels improves performance also holds for the other datasets we studied.

It is less surprising that MTLSL also outperformed gold training according to all four soft evaluation metrics: It produces a distribution less divergent from that of the annotators (as measured using cross entropy and Jensen-Shannon divergence) and better captures item confusion (as measured by cosine similarity and Pearson correlation). Of the other approaches to using crowd information to supplement gold labels, MTLOA and PEWE, MTLOA always outperformed gold training, sometimes significantly so, but PEWE fell behind gold training, significantly so when evaluated using cross entropy and cosine similarity.

Finally, the fact that the MTLSL method for supplementing gold labels was more effective than MTLOA and PEWE suggests that, with this dataset, targeting the distribution of labels is more useful than targeting observed agreement or confusion among labels.[27]

7.2.3 Learning from Aggregated Labels

With hard evaluation, training on this dataset using aggregated labels produced results significantly worse than training from gold labels and slightly worse than training from soft labels. With soft evaluation, training with aggregated labels yielded worse or significantly worse results than training with soft labels but better or substantially better results than training with gold labels. Majority voting (MV), Dawid and Skene (D&S), and MACE achieved comparable results according to all metrics.

As we will see, this result was replicated with the other datasets: Training against the entire "soft label" yielded as good or better results than training against aggregated silver labels with all of our datasets and with all types of evaluation. This suggests that the distribution of labels produced by the annotators generally does provide useful information, which is lost when the soft label is aggregated. Training with aggregated labels only matches training with soft labels with datasets such as PDIS, where (relatively) low "coder accuracy" is actually the result of a lot of noise in the annotations (as opposed to ambiguity as in POS), yet there is an abundance of annotations per item, allowing an aggregation method to learn accurate models. This results in high-quality aggregated labels for PDIS, much higher than obtained with MV. For most of the evaluation metrics, aggregation methods did not outperform MV on POS; these methods also performed about the same with hard evaluation as the DLC method (Rodrigues and Pereira, 2018), which likewise attempts to learn coder models and uses them to weight interpretations. We take these results as further evidence that in POS, a lot of disagreement is informative.

7.2.4 Learning from Soft Labels

The best results for this dataset without using gold labels for training were obtained by two soft-loss methods, KL and CE, using as target the distribution obtained by softmaxing the raw proportions, rather than standard normalization or the output of probabilistic

27. It should however be noted that PEWE outperformed gold training in the original paper (Plank et al., 2014a), in which the method used the confusion among *two* expert annotators. Perhaps a better way exists to extend PEWE to a multi-annotator scenario than the one used here.

aggregation. The best results were obtained by using KL as a soft loss function; training using CE as a soft loss function achieved slightly worse but not significantly different results. These two methods also obtained the best results according to soft evaluation metrics, both in producing the probability distribution least divergent from the annotators' label distribution and for capturing confusion as measured by entropy. These results provide further evidence that the crowd information included with this dataset provides useful information for learning; the fact that the entropy of the soft-label distributions is highly predictable is particularly significant in this respect.

The next best results were obtained by a cluster of methods including MSE soft loss, SREL, and the soft approximation of the CrowdTruth aggregation method (see Section 5.2.1). These methods were significantly outperformed by the KL and CE soft-loss methods, but significantly outperformed the other soft-label method, DLC. All soft-labelling methods outperformed training using aggregated labels.

7.2.5 Filtering and Weighting by Item Difficulty

Using crowd information (specifically, observed agreement on an item) to filter "hard" items generally resulted in significantly worse hard evaluation performance:[28] Training with gold labels with hard items filtered always resulted in worse performance than training with all the gold labels, and training with MV silver labels with hard items filtered always resulted in worse performance than training with the entire dataset with one exception discussed below, IC-LABELME. POS is a clear illustration of this trend. For F1 evaluation for example, filtering hard items before training on gold labels resulted in a drop of 12 F1 points below gold training without pre-filtering, and training on labels aggregated using majority voting after filtering hard items fell 8 F1 points below training using MV silver labels without pre-filtering. However, with soft evaluation (with which hard label methods generally perform worse anyway), the effect of filtering was less clear-cut: In some cases we saw an improvement, in others we did not. With POS, training on gold labels after filtering hard items (gold + OA filter) led to significantly better soft evaluation results than training on gold labels without pre-filtering; however, the reverse was the case for MV + OA filter and MV training.

Augmenting silver (MV) labels by weighting the loss of each item according to some measure of confusion, such as the observed agreement for that item (OA weighting), generally worked better than filtering for hard evaluation. Besides again yielding much better results with IC-LABELME, it also achieved better results than training with unweighted MV labels on MRE, RTE, and IC-CIFAR10H, although the difference was generally not significant. But weighting did not typically affect soft evaluation results. With POS we did not see an improvement over using MV labels alone according to either hard or soft metrics; in fact we found significantly worse cosine similarity and cross entropy results. This may suggest that, given the complexity of the annotations discussed above, OA alone is not informative about the nature of disagreements for this task.

28. We remind our readers that the item difficulty models we are aware of, such as GLAD (Whitehill et al., 2009), are not applicable to this dataset as it is not binary.

7.3 Information Status Classification

In the PDIS dataset, gold labels are only available for testing, not for training, so it is not possible to report results for training with gold labels, and our discussion will focus on the results obtained with aggregated labels and soft labels. We did evaluate the performance of "hybrid" methods relying both on a hard label and information from soft labels PEWE, MTLSL) on this dataset. However, our hard label was the most accurate aggregated label (we used the labels produced by MACE, but D&S was just as accurate) instead of a gold label as in other datasets. The results with "gold" and "augmented gold" labels are therefore not directly comparable to those obtained with other datasets.

The key characteristics of PDIS (see Tables 1 and 2) are that it has the highest number of items (96,305), and the average number of annotations per item is also fairly high (11.87). Observed agreement is medium high (.81). However, average coder accuracy is mediocre (.78), and the percentage of "expert" coders is low (.71). Notwithstanding this, the quality of aggregated labels is high, .98. Finally, the entropy statistics are the opposite of those obtained with POS: while the raw annotator entropy is fairly high (.38), the best distribution entropy (BDE) is one of the lowest (.09).

7.3.1 Using Soft Labels to Supplement (Silver) Hard Labels

As with POS, MTLSL was the best method for learning hard or weighted truth for PDIS under all three hard evaluation metrics; unlike with POS, however, the improvement over only using the (silver) hard label was not significant. MTLOA and PEWE also performed on par with MACE. The explanation proposed when discussing POS in the previous subsection—that soft labels provide information that can lead to improvements over the hard labels when the dataset contains sufficient quality, quantity, and diversity in the soft labels—can be applied to explain the results with PDIS as well. In the case of PDIS, while we have a high number of crowd judgments, their quality is lower than with POS (average accuracy is .78; percentage of "expert" coders is .71), and above all we have much less diversity, as measured by our best distribution entropy measure (which in this case was obtained from the posterior probability): .09, as opposed to .39 with POS. This substantial difference suggests that many more of the coders' disagreements with regard to the gold labels are due to errors or other noise, and thus can be identified by a probabilistic aggregation method, in comparison with POS, where most of disagreements are due to ambiguity or overlap, and thus provide useful information.[29] This hypothesis is confirmed by analyses of the disagreements such as those reported in Section 2.8.1. We would therefore expect a smaller improvement over only using the hard labels, as indeed was the case.

With soft evaluation, augmenting MACE with crowd information generally improved results. In particular, MTLOA and PEWE outperformed MACE and in fact all the aggregated- and soft-labelling methods according to three out of the four evaluation metrics (all except Jensen-Shannon divergence). However, MACE outperformed MTLSL using all evaluation metrics except Jensen-Shannon divergence. This suggests that information about label and/or annotator confusion is more useful for this dataset than the probabilistic output of the MACE aggregation model used as a soft target for MTLSL (see Section 7.1). This provides

29. Note also that whereas ambiguity is possible, indeed frequent, in this dataset, label overlap isn't: either an entity has been previously mentioned, or it hasn't.

further evidence for the hypothesis that much of the disagreement in this dataset is due to noise.

7.3.2 Learning from Aggregated Labels

As shown in Table 1, PDIS is a very mixed dataset in terms of annotator performance. We have already discussed how average coder accuracy is not very high, 78%, and the variation is much wider than in the other datasets. In addition, the annotators did varying amounts of work, annotating from about 1% to 13% of the dataset; the majority of the annotations, however, were produced by the annotators doing the most work. We would therefore expect, first of all, probabilistic aggregation methods to perform much better than majority voting with regard to hard evaluation, as MV's assumption that all annotators have similar ability clearly does not hold. Meanwhile, unlike with POS, probabilistic aggregation methods have a lot of evidence from which to learn accurate characterizations of the annotators that produced most of the labels. This prediction is borne out, first, by the fact that the quality of probabilistically aggregated labels (98%) is much higher than the quality of MV labels (89%) (see Table 2) and the quality of aggregated labels in POS (at most 80%). And second, by the fact that training with probabilistically aggregated labels outscored training with MV labels by at least 2 percentage points with all three hard evaluation metrics (see Tables 7, 8, and 9).

A second prediction is that having access to the entire distribution of labels produced by the crowd should be less informative in terms of predicting the most likely label for this dataset than for POS, primarily because the diversity of labels as estimated by the best distribution entropy is so low (.09) but also because the quality of coders as estimated in terms of observed agreement is much lower and coder ability is highly variable. (One could also think that given that probabilistic aggregation methods achieve such high accuracy it would be difficult to improve on them, but this is not the case e.g., with IC-CIFAR10H, as we will see.) And indeed, for this dataset, training using aggregated labels performed on par with training using soft-labelling methods.

7.3.3 Learning from Soft Labels

Soft labels do not appear to be entirely uninformative, however. As we stated earlier, with MTLSL, making use of both the hard aggregated silver and the soft label outperformed training with aggregated silver only, even if only marginally. More importantly, soft-labelling methods and/or versions of MTL outperformed pure aggregated label training with all soft evaluation metrics.

While the performance of aggregated labels was on par with augmented methods and most soft-labelling methods (the CE/KL/MSE soft-loss methods) with hard evaluation, these soft-loss methods outperformed training with aggregated labels when it came to learning the entropy of the annotator-produced distributions, a proxy for labelling uncertainty that we measured using cosine similarity and the Pearson correlation of the entropy.

However, a striking feature of this dataset is that all methods were quite bad at predicting entropy. This is also the only dataset for which the best distribution was obtained from the posterior of a probabilistic aggregation method (MACE) rather than directly from raw annotations using softmax or standard normalization. Again, these two findings suggest

that crowd information for this dataset is very noisy, so no method can learn to predict soft labels accurately. The second finding likely also explains why none of the soft-loss methods (CE/KL/MSE) improved even insignificantly over their hard-label counterpart for this dataset: The soft labels obtained via MACE have had too much disagreement information removed to be useful.

As mentioned above, the CE/KL/MSE soft-loss methods were on par with the aggregated (and augmented) methods when it came to hard evaluation metrics. These methods were also on par with DLC when evaluating using accuracy and F1 and slightly outperformed it when evaluating using CT F1. Like the MACE and D&S aggregation models, DLC learns ground truth by learning annotator reliability. But the CE/KL/MSE soft-loss methods and DLC all significantly outperformed SREL, which is based on raw coder annotations. This again shows that for PDIS, gains in hard evaluation performance are seen with models that discriminate between annotators/annotations. Further evidence is the fact that although the CT soft aggregation method outperformed MV, it was the least performant soft-label method when it came to hard evaluation. While the best of the CE/KL/MSE methods always outperformed the best aggregated method when evaluated using soft metrics, the same cannot be said of DLC, SREL, and the CT soft-aggregation method, as we will see when discussing other datasets.

7.3.4 Filtering and Weighting by Item Difficulty

As with POS, pre-filtering then training resulted in lower performance than training on MV labels alone when evaluated using all the metrics. While we did not have gold labels for training, we observed that pre-filtering and training using the very high quality MACE aggregated labels also led to a worse performance according to all the metrics. This is likely because with such a noisy dataset, a lot of items would be eliminated by filtering.

While weighting using observed agreement or using the item difficulty scores produced by the Whitehill et al. (2009) method outperformed training using MV alone when using soft metrics, it did not lead to a higher hard evaluation performance. Another interesting finding is that weighting with probabilistically inferred inverse difficulty generally results in worse performance than weighting with OA.

7.4 (Medical) Relation Extraction

The key characteristics of MRE (see Tables 1 and 2) are that it is one of the smallest datasets we studied, with only 975 items, and coder accuracy against gold labels is also fairly low: Average coder accuracy is even lower than with PDIS (.76), and the percentage of "expert" coders is lower still (.58), although observed agreement among annotators is reasonably high (.86). The average number of annotations per item is fairly high (15.3), but the quality of aggregated labels is the lowest among all of our datasets (highest is .77). Both raw annotator entropy and best distribution entropy are fairly high (.31).

Another interesting observation is that, going by the gold label, the dataset is very imbalanced, with a nearly 3:1 ratio between class 0 ($false$) and class 1 ($true$). As a result, the accuracy ranking for various methods often differed from their F1 or CT F1 ranking. Because we used the class-weighted version of the F1 metric, we expected results would differ, as the metric would assign a higher score to the model that produced more correct

answers of class 0. And, on this note, a striking result is that if the goal was to learn the majority class (i.e., evaluated using F1 or the weighted F1 metric), the CrowdTruth method outperformed all other methods for learning from crowds, confirming the results obtained by Dumitrache et al. (2018a).

7.4.1 Gold vs. Non-Gold

With MRE, as with POS, we found a large difference in performance between the results obtained when training from the crowd only and training using gold labels when measured using hard evaluation metrics. We observed a similar difference with POS (Section 7.2), for which we indicated that the most likely explanation was low coder accuracy against gold labels for nearly all classes except nouns and pronouns. The same explanation applies to MRE, as we can see without digging into the dataset: MRE and IC-LABELME, discussed in Section 7.6, are the datasets with the lowest average coder accuracy (measured against gold labels) and the lowest proportion of "good" annotators (again, measured against gold). In other words, these are the datasets where the crowd produced labels least like the gold. As seen in Section 2.8, perhaps the majority of disagreements from gold in POS are due to ambiguity, although there is some label overlap, whereas in IC-LABELME, label overlap seems to play a substantial role. MRE would appear to be like IC-LABELME in this respect (see also the discussion in Dumitrache et al. (2018a) and Dumitrache (2019)). It is therefore not surprising that the models trained on these soft labels alone also produce labels that substantially differ from the gold labels.

7.4.2 Using Soft Labels to Supplement Gold Labels

As noted above, using gold labels in training yielded the best results for MRE with hard evaluation. The best results with all hard metrics were obtained by supplementing gold labels with soft labels, but the improvement over using only gold was typically not significant. (MTLSL worked slightly better according to accuracy and CT F1, PEWE according to F1.) The fact that we saw a small but not significant improvement is, we believe, consistent with the hypothesis proposed in Section 7.2.2 about the conditions under which an improvement can be seen: MRE has a fairly high BDE, indicative of a good level of diversity, but not as high as that of POS; it has a good number of annotations per item, but the size of the dataset is likely too small to observe an effect, and coder accuracy is also fairly low.

With soft evaluation, we found that one of the gold + soft label methods achieved slightly better results than training with gold only according to CE or JSD, but slightly worse when measured using cosine similarity and entropy correlation. The most likely explanation is that in the MRE dataset the annotators often have to choose a label against many applicable ones, so the entropy may not be predictable; but it may also be a matter of size.

7.4.3 Learning from Aggregated Labels

Several interesting observations can be made on the basis of the hard evaluation tables in Section 6.1. First of all, we can see that unlike with POS, where soft-label training generally outperformed training with aggregated labels, pretty much all crowd-only training methods achieved about the same results in terms of accuracy (Table 7), although some interesting differences can be seen with F1 and CT F1. Second, we found that there was a clear winner

for this dataset among the crowd-only methods in terms of F1 and CT F1: CrowdTruth aggregation, which achieved a performance almost 10 points higher than any other method. As CrowdTruth is best considered a soft-label method, we discuss this finding next. Third, we found that, again, the comparison between probabilistic aggregation methods and MV was very much affected by the hard evaluation metrics. With accuracy, all aggregation methods performed about the same, as previously stated. With F1 and CT F1, however, D&S performed much better than both MV and MACE in terms of both hard and soft evaluation—this is the only dataset for which we found a substantial difference between D&S and MACE under either form of evaluation. Together with the finding about the performance of CT aggregation, this result suggests that D&S is better than either MV or MACE at modelling the main class.

With soft evaluation, silver training with D&S outperformed gold training both in learning the distribution of the annotations (i.e., evaluation using JSD and CE) and according to the entropy similarity metrics but was outperformed by soft-labelling methods.

7.4.4 Learning from Soft Labels

The difference in F1 and CT F1 performance between CT "aggregation" and all other crowd-only methods with this dataset is, we believe, due to the same reason that explains the better performance of D&S over MACE aggregation: the focus on the true class. D&S aggregation learns models of the coders' sensitivity and specificity to the true class; likewise, the objective of CT aggregation is to find good examples for the true class.

However, other soft-label methods apart from CT aggregation did not improve results over silver training when evaluating using hard metrics. This is most likely due to the fact that this dataset does not satisfy any of the conditions under which soft-label methods achieve good performance: It is the second smallest, and the quality of the annotations is the second lowest. However, soft-label training did outperform hard-label (silver and gold and augmented) training when evaluating using soft evaluation metrics.

7.4.5 Filtering and eighting by Item Difficulty

One clear result for filtering and weighting on this task is that both approaches led to significantly worse accuracy than non-filtering/non-weighting using hard evaluation metrics. And, while filtering led to better soft evaluation results, weighting largely remained on par with non-weighting. Neither method led to gold-level hard evaluation performance.

7.5 Recognizing Textual Entailment

The key characteristics of the RTE dataset are that it is the smallest dataset, counting only about 800 items, but it has a good number of annotations per item, 10. The extent of agreement between the coders and gold labels, as measured by average coder accuracy (0.84) and percentage of expert coders (0.83), is quite good, although not as high as that of IC-CIFAR10H. The average number of annotations per coder is not, however, at 48.8, very high, and the observed agreement between coders, 0.63, is also quite low—in fact, it is the lowest among all datasets.

This low agreement between coders appears to be due primarily to the fact that many items in the RTE dataset are "difficult" in the sense that it is unclear whether the premise

entails the hypothesis or not, as discussed in Section 2.8.4. The fact that agreement with gold labels is quite high suggests that the difficult items are not the majority, which would confirm the findings of, e.g., Pavlick and Kwiatkowski (2019)—in their data, about 20% of cases are difficult.

7.5.1 Gold vs. Non-gold

One obvious characteristic of RTE is that although using gold labels still yielded the best hard evaluation results (with gold or gold + soft achieving the best results depending on the metric), the margin between training with gold labels and training with crowd labels was minimal, much smaller than with the two datasets we have seen thus far, POS and MRE; in fact, soft-loss and weighting methods achieved equivalent results to using gold labels for accuracy and F1. (A direct comparison to PDIS is not possible, as that dataset has no gold labels for training.) This result was previously reported by the creators of this dataset, Snow et al. (2008), but without explanation. We believe it can be accounted for with reference to coder accuracy. In RTE, the coders have a much higher average accuracy with respect to gold labels, and the percentage of expert coders is very much higher, than in MRE in particular. As for POS, as discussed in Section 7.2.1, the headline coder accuracy and expert percentage figures are deceptive, in that accuracy is only high with one category, while for the others it is quite low.

Further evidence for this explanation is the fact that the quality of aggregated labels is very high even though each annotator only produced relatively few annotations. The majority voting accuracy is already 90% with respect to gold (or 93% depending on how the ties are broken). The other hard aggregation methods also produced labels with 93% accuracy. It is therefore not surprising that the performance margin between gold training and crowd-based training was virtually nil.

In a nutshell, we are arguing that with this dataset, the gold labels are not different in quality from the labels provided by the crowd. On the other end, unlike with IC-CIFAR10H in particular, there are grounds for disagreement, namely, difficulty, which explains why the agreement between coders is low (as well as, we would argue, the substantially lower values reached for hard evaluation compared to the other NLP datasets, and in particular to MRE, which has a comparable size).

7.5.2 Using Soft Labels to Supplement Gold Labels

With RTE, gold-plus methods did not result in significant improvements in hard evaluation over gold-only training, unlike with POS or MRE. Supplementing the gold labels with crowd information led to non-significant improvements over gold training in terms of accuracy, to equivalent results with the other metrics (slightly lower, but again, the difference was not significant). In our view, this is in part due to the fact that the quantity of data is much smaller than with POS, in part to the fact that the diversity of the labels, as measured with BDE, is lower than with MRE (and half that of the diversity we see with POS, where gold + soft–methods did significantly improve over gold). As for soft evaluation, we found that MTLSL achieved significant improvements over gold in terms of JSD, but otherwise the results obtained supplementing gold labels with crowd information were comparable to those obtained training with gold labels alone.

7.5.3 Learning from Aggregated Labels

Training using any silver label achieved slightly lower results with RTE than training with gold, less than one accuracy/F1/CT F1 point—a margin that is significant but much lower than that observed with POS or MRE. This might seem surprising given that the number of annotations per coder is relatively small, but we think it can also be explained as the result of the high quality of the crowd annotations, making them essentially comparable to gold annotation. This hypothesis is confirmed by the fact that MV achieved comparable results to the probabilistic aggregation methods. None of the silver-label methods was significantly outperformed by any other method for learning only from crowds. For soft evaluation, in most cases, there was no significant difference between training using gold labels and training using silver labels: Both gold- and silver-label training methods were outperformed by soft-labelling and augmented methods.

7.5.4 Learning from Soft Labels

When discussing the results with POS in Section 7.2.3, we pointed out how soft-label training achieved as good or better results in terms of hard evaluation than aggregate labels with all datasets. Specifically, there are three datasets with which soft-label training gave better results—POS, IC-LABELME, and IC-CIFAR10H—and three with which the results were equivalent—PDIS, MRE, and RTE. What characteristics do these last three datasets have in common?

In Section 7.2.3, we argued that training with aggregated labels matches performance with soft labels when average coder accuracy is relatively low but there are enough annotations per item and per coder to allow the aggregation method to acquire good models of the coders, resulting in high-quality aggregated labels. We saw in Section 2.2 that these conditions held for PDIS; they hold for RTE as well. They do not, however, hold for MRE. But there are two additional characteristics common to these three datasets. The first is that these three datasets, for which soft-label training does not improve over silver aggregate training, are all binary classification tasks. It may be that in terms of hard evaluation, a model trained for binary tasks is always better off "taking a stand" as opposed to taking a probabilistic approach to truth. Another characteristic these datasets have in common is that they have the highest raw distribution entropy (see Table 1). Soft-loss training is, perhaps, not especially tolerant of confusion.

More specifically, it is the soft-loss methods that performed on par with silver-training methods on this dataset, outperforming repeated labelling and DLC. In fact, the repeated-labelling method SREL achieved worse results on this dataset than training using MV labels with all hard metrics—this is the only dataset for which this happened. RTE is also the dataset with the highest item entropy (0.72, 0.34 points higher than the next highest, PDIS). This is consistent with what we remarked before about the high level of difficulty-induced disagreement in this dataset. Taken together, these facts suggest that SREL is not suited for datasets with such characteristics. This hypothesis is further strengthened by the fact that the next method for which repeated labelling achieved much worse results than the best silver or soft-loss method is PDIS, the dataset with the next highest entropy.

With soft evaluation, the results were somewhat mixed. However, we can definitively say that, with the exception of Jensen-Shannon divergece, the soft evaluation methods achieved the best results.

7.5.5 Filtering and Weighting by Item Difficulty

As with the datasets seen thus far, filtering items with low agreement did not yield any improvements over training using all the items in terms of hard evaluation, and weighting by observed agreement did not achieve better results than using majority voting labels without weighting items. However, unlike what we observed with PDIS and with MRE, weighting using the inverse difficulty scores inferred by the Whitehill et al. (2009) aggregation model resulted in a substantially worse performance when evaluated using hard metrics.

7.6 Image Classification 1: LabelMe

IC-LABELME's most distinctive features are the low number of annotations per item (2.5 on average), the extremely low coder accuracy with respect to gold (.69 average accuracy, with only 42% of coders achieving expert accuracy levels), and the extremely high BDE (.76, almost double the next highest).

7.6.1 Gold vs. Non-gold

With IC-LABELME we again found a large difference between training using gold labels and training using only crowd information. With hard evaluation, we found that using gold labels resulted in an advantage over soft-label training of more than 10 percentage points for accuracy and F1 and slightly less for CT F1, similar to what we observed with MRE and POS. The same explanation we proposed for MRE and, after some analysis, for POS—that the reason for the large difference is the substantial difference between coder judgments and gold judgments—applies to IC-LABELME as well: These are the datasets with the lowest coder accuracy and the lowest percentage of expert-quality annotators when evaluated against gold annotations.

By contrast, with soft evaluation, the situation was exactly reversed, and we found a large difference with all soft evaluation metrics in favor of methods using crowd information, either by itself in soft-labelling methods—in particular, soft-loss methods, but also training with aggregated labels—or in combination with gold labels. MTLSL performed on par with soft-loss methods when measured by cross-entropy and JSD and only slightly worse in terms of the entropy similarity measures.

This reversal confirms what was already obvious from the example discussed in Section 1: that gold judgments are very different from crowd judgments in this dataset. However, as discussed in Section 2.8.2, with IC-LABELME the discrepancy between gold and crowd judgments appears to be due to the degree of overlap among the labels in the annotation scheme, rather than ambiguity as in the case of POS or carelessness as in the case of PDIS. Further evidence for this is the extremely high BDE, by far the highest in any of the datasets we used.

7.6.2 Using Soft Labels to Supplement Gold Labels

As already mentioned in discussing gold vs. non-gold, very different results were achieved with this dataset by leveraging crowd information in addition to gold labels depending on which form of evaluation was used.

The best hard evaluation results against gold for this task were obtained by training with gold labels alone: Supplementing gold labels with crowd information led to a lower performance than training with gold alone, which is significant in all cases except with F1 evaluation of MTLSL; the difference between gold only and augmented gold with MTLSL is small. The simplest explanation for this would be that given the high level of randomness in the choice of labels, training using gold labels is the best way to optimize for testing against gold labels.

But whereas crowd information did not improve upon gold labels for learning hard truth, MTLSL always significantly outperformed gold-only training with soft evaluation, as did the other gold-plus training—the exception being the entropy correlation results, where PEWE and MTLOA-only remained on par with gold training. (Gold-plus methods were, however, generally outperformed by soft-loss methods with soft evaluation, except again for MTLSL, which achieved equal-best performance with the soft-loss methods in terms of cross-entropy and JSD and near-best with the entropy correlation metrics.)

There are at least two reasons for this difference in results. First of all, as already discussed, although training with crowd information alone can match or indeed outperform gold training, this only happens when certain conditions are met, which is not the case with IC-LABELME. With IC-LABELME, the average number of annotations per item is only 2.5, with a maximum of 3, and over 4% of the items only have a single annotation. In other words, the number of annotators per item is insufficient, meaning the crowd annotations do not contain additional information for gold augmentation/regularization. And second, crowd judgments are very different from gold judgments with this dataset, as already noted above. As a result, methods relying on one type of judgment generally performed badly when evaluating against the other type, and vice versa—the one exception being MTLSL, which optimizes for both.

7.6.3 Learning from Aggregated Labels

With this dataset, as with PDIS, probabilistic aggregation methods outperformed majority-vote aggregation by a substantial margin when evaluated against the gold label, and for the same reason: the low similarity between coder judgments and the gold labels, or between the judgments of different coders. But while training with probabilistically aggregated labels outperformed MV, all silver methods were outperformed by soft-labelling, weighting, and filtering methods using all the evaluation metrics.

7.6.4 Learning from Soft Labels

Soft-loss training significantly outperformed all other soft-labelling methods with IC-LABELME in terms of hard evaluation, except for weighting and filtering (see next subsection). Soft-loss methods also produced the best soft evaluation results with almost all metrics.

7.6.5 Filtering and Weighting by Item Agreement

ic-labelme is the only dataset for which filtering items by observed agreement, and then training over the remaining items, resulted in an improvement over training without pre-filtering. In fact, with this dataset, filtering + mv labels was the best approach to learning from crowds. We believe that this is again because this dataset is the one in which annotators disagree the most with the gold labels, as shown by the low coder accuracy figures—and also by the fact this is the only dataset in which the expert annotators do not constitute a majority in the annotator population. Because the base model for this task was pre-trained with previously learned and encoded images, the model loses nothing by discarding low observed agreement and perhaps mislabelled items.

7.7 Image Classification 2: CIFAR-10H

ic-cifar10h is not a particularly large dataset—it is comparable to pos or ic-labelme—but it has the highest number of annotations per item, over 50. It also has very high annotator accuracy, with all annotators having an accuracy of 75% or more. The only other dataset with a percentage of coders with a similarly high degree of agreement with the gold label is pos; but, unlike in pos, in ic-cifar10h the annotators did not overwhelmingly label only one category. Also, each coder annotated about 200 items on average. As a result of high number of annotations per item, coder quality, and a good number of items being annotated per annotator, the quality of the aggregated labels is the highest, .99, for this dataset, irrespective of the type of aggregation used. Finally, this is a dataset with very high oa and very low entropy, both raw and bde.

These last characteristics, as well as the high coder accuracy, appear to be due to the fact that the image categories are clearly distinct, unlike in ic-labelme, and that there is no ambiguity. However, some items are more difficult to label because of blurriness and small size, as discussed in Section 2.8.4.

7.7.1 Gold vs. Non-gold

Another result of the high similarity between gold labels and crowd labels and the high number of annotations for ic-cifar10h is that it is the one dataset for which training with crowd information outperformed training with gold labels, regardless of the method used. We already mentioned in connection with rte the finding in Sheng et al. (2008) and Snow et al. (2008) that a large enough crowd may produce labels of quality comparable to that of gold labels produced by experts when the crowd workers are of sufficient quality; ic-cifar10h shows that in fact the crowd can outperform experts.

7.7.2 Using Soft Labels to Supplement Gold Labels

Two out the three methods for augmenting gold labels, pewe and mtlsl, resulted in significantly reduced hard evaluation performance with respect to gold on this dataset; only mtloa achieved a performance on par with gold training. We saw the same result with ic-labelme, and again we hypothesize that the reason is that the crowd annotations do not provide useful additional information for gold augmentation/regularization. However, here, the reason is different. With ic-labelme, the motivation for the lack of improve-

ment was the low number of annotations and the low level of agreement of the annotators compared to gold. For IC-CIFAR10H, however, the reason is that the crowd annotations do not provide enough diversity compared to gold labels, as they appear to be drawn from the same distribution: There is hardly any disagreement between gold labels and soft labels. This can be seen from the combination of high accuracy and high observed agreement of the crowd labels with respect to the gold. We can also see that both the raw annotation entropy and the BDE are extremely low, the lowest among all the datasets. Further evidence is that the gold + soft methods did not even outperform gold training in terms of soft evaluation—again, the only dataset for which this was the case.

7.7.3 Learning with Aggregated Labels

D&S and MACE did not significantly improve over MV for this dataset, regardless of the evaluation metric. This is unsurprising given the quality of the coders—labels aggregated using majority voting already achieved the same accuracy (over 99% with respect to gold labels) as probabilistically aggregated labels, a sign that discriminating between annotators would offer little improvement over the majority with respect to learning ground truth. For the same reason, silver training was not significantly distinguishable from gold training.

7.7.4 Learning with Soft Labels

As already discussed, IC-CIFAR10H demonstrates that a large crowd providing high quality annotations can not only match, but outperform gold training. For this task, the soft-labelling methods outperformed all types of hard-label training, both silver and gold. Among soft-labelling methods, repeated labelling and DLC outperformed soft-loss and aggregated methods according to all the hard evaluation metrics, but soft-loss methods still outperformed all hard-label methods, both gold and silver. The results with soft evaluation were more complex: Repeated labelling and soft-loss training achieved the best results with cross-entropy by a wide margin over DLC, which, however, outperformed all other methods when evaluated using Jensen-Shannon divergence. Soft-loss methods achieved the best results in terms of entropy estimation.

7.7.5 Filtering and Weighting by Item Difficulty

For this task, training with MV labels but weighting the loss for each item depending on the observed agreement for that item led to a small improvement over majority-voting training. This only happened with one other dataset: IC-LABELME. In the case of IC-LABELME, one could argue that the dataset contains lots of difficult items, as shown by the low overall agreement, and that observed agreement works well at identifying such items. However, the reason why weighting improved results over MV for IC-CIFAR10H as well (whereas filtering does not) is not immediately apparent. But the effect is small, and at any rate, MV training is not the highest-performance method.

8. Discussion

In this section, we address the research questions put forth at the beginning of this work, and discuss the results as answers to these questions.

8.1 Are gold labels required to achieve the best results in training (RQ3a)? And does the answer depend on the form of evaluation (RQ4a) and dataset (RQ4b)?

Our clearest (if possibly least surprising) result is that the answer to **RQ4a** (but also to **RQ4b**) in the context of **RQ3a** is very much the affirmative. Training with gold labels as the target almost always achieves the best results when the evaluation is against gold labels, although this depends on the characteristics of the dataset (see below). But with soft evaluation against soft labels—aka probability distributions extracted from the annotators' judgments—the best results are always achieved by only using these soft labels as targets in training, irrespective of the dataset, although methods leveraging both gold and soft labels generally match, if never exceed, the performance of soft-label training methods. This finding may seem obvious in retrospect, but as far as we know it has not been previously discussed in the literature.

Furthermore, our results with IC-CIFAR10H, in particular, indicate that under certain conditions, models trained without assuming a gold truth can achieve better performance than models that leverage gold labels. Among our datasets, there were three with which using gold resulted in substantially better performance with hard evaluation (POS, MRE, IC-LABELME), and one with which the difference, while significant, was minimal (RTE). But with IC-CIFAR10H, training using soft labels without gold worked better than training with gold. (We did not have gold for training with one dataset, PDIS.) Our results reveal that when a dataset is sufficiently large—and is annotated by a large number of high-quality coders according to an annotation scheme such that annotators frequently agree with each other—a number of ambiguity-aware training methods produce better results than training on gold labels. IC-CIFAR10H met these conditions, and RTE got close. We also noted that POS is an apparent counterexample to this hypothesis, given that gold training outperformed training with only crowd information by a large margin even though the dataset has high coder accuracy and a substantial number of annotations per item. But our analyses indicate that the high overall coder accuracy and the high overall number of annotations for POS are in fact deceptive, as the dataset is divided in two subsets: one with lots of annotations and high accuracy and one with a very low number of annotations and low accuracy. And the high accuracy is mostly for nouns, whereas with other part-of-speech tags the accuracy is much lower.

8.2 Does Supplementing Gold Labels with Crowd information Result in Better Performance than Training with Gold Only (RQ2a)? And Does the Answer Depend on the Form of Evaluation (RQ4a) and the Dataset (RQ4b)?

The answer to **RQ2a** also very much depends on the form of evaluation and the dataset. With soft evaluation, leveraging crowd information in addition to gold labels results in better or significantly better performance with almost all forms of evaluation and almost all datasets; the only exceptions are RTE with cross-entropy and IC-CIFAR10H with JSD.

With hard evaluation, there are three datasets in which supplementing gold information with information from the crowd (leveraging the soft label) helped: significantly with POS (1 p.p. gain), marginally with PDIS and MRE. With RTE, again there was no significant

difference, and which method achieved better performance depended on the metric. But with IC-LABELME and IC-CIFAR10H, using soft labels in addition to gold hurt performance. This last point is particularly surprising in light of the fact that with IC-CIFAR10H using only crowd information achieved much better results than using gold.

Our proposed explanation for these hard evaluation results is that soft labels help gold-label training "when the soft label provides useful information beyond the preferred label that leads to a better model"—i.e., when the soft label helps regularize gold-label training. In order for this to happen, two conditions must hold. First of all, the decision on the best label for an item must be sufficiently complex, on average. We propose that this complexity can be measured using average best distribution entropy (BDE): If the BDE is too low, leveraging soft labels in addition to hard labels does not help—which is why MTLSL outperforms gold with POS, but not with IC-CIFAR10H, for which the BDE is nearly 0, even though coder accuracy with IC-CIFAR10H in particular is very high. In other words, where the soft labels mostly reproduce the gold standard, their informative contribution lessens. Second, there have to be enough judgments for the soft label to be sufficiently reliable. This explains why we only see marginal improvements with IC-LABELME (too few annotations per item) and RTE (too few items).

8.3 Which Method for Learning from Disagreement Achieves the Best Results (RQ3b)? And Again, to which Extent Does the Answer Depend on the Form of Evaluation (RQ4a) and the Dataset (RQ4b)?

One of the key results of this paper is that the answer to **RQ3b** is more complex than one would expect based on the previous literature. Two points are, however, very clear. First of all, soft-label training generally outperforms aggregate- (silver-) label training with all datasets and all forms of evaluation. And second, which soft-labelling method performs best very much depends on the form of evaluation and the characteristics of the dataset.

With soft evaluation, some form of soft-label training achieved the best results with virtually all datasets and all metrics, except with Pearson correlation of entropy with RTE, the smallest dataset. Specifically, some form of soft-loss training achieved the best results with all datasets except for RTE and MRE; repeated labelling achieved the best results with MRE (all metrics) and RTE (cosine similary of entropy); and deep learning from crowds generally achieves worse results than the other soft-label methods, except with RTE (CE) and IC-CIFAR10H (JSD).

With hard evaluation, which method performed best very much depended on the characteristics of the dataset. The most substantial gap in performance was observed with MRE, a dataset focusing on positive examples, and CrowdTruth "aggregation"—more of a soft label method in our extension to multiple classes—achieved substantially better F1 and CT F1 results than any other method. Soft-loss training achieved competitive results for all the multi-class datasets for which annotator distribution has a low average entropy (POS, IC-LABELME, and IC-CIFAR10H), but it was the best method for only POS, a dataset for which the average annotator (gold) accuracy is high but the accuracy of aggregated labels is unexpectedly low. With IC-LABELME, which features a low level of alignment between gold labels and most annotators (perhaps owing to the high subjectivity of the task and arbitrariness of the gold truth—see Section 1), the best method was pre-filtering low OA

items and training on the rest. With IC-CIFAR10H, characterized by a high number of per-item annotations produced by high-quality coders, SREL and DLC, both of which multiply examples, outperformed other methods. Probabilistic aggregation techniques such as D&S and MACE worked best for PDIS, which has a large number of coders of varying ability.

8.4 How Can We Best Evaluate Models on Datasets that Provide a Range of Judgments (RQ1)?

And finally, we return to **RQ1**. Throughout this discussion we have highlighted how much the relative ranking of the methods for learning with disagreement depends on the form of evaluation. The inevitable conclusion is that evaluating models using only hard evaluation metrics such as accuracy or F1, or only using soft metrics such as cross-entropy, will only provide a partial picture of how well a model performs on a dataset. A hard evaluation metric is only truly appropriate for datasets on the low disagreement end of the spectrum, as indicated, e.g., by what we have called best distribution entropy, or BDE. In all other cases, also reporting the results with a soft evaluation metric is arguably more accurate.[30]

On the other end, one could ask whether all of the metrics we studied in this paper are required. With regards to hard evaluation, we can see that accuracy, F1, and CT F1 rank the learning methods similarly, except on the MRE dataset, which is highly imbalanced (see Section 7.4). Also, while F1 and CT F1 ranked the methods in a very similar way, we can see that the CT F1 metric increased the scores of all the methods. Comparing the F1 and CT F1 scores in Tables 8 and 9,[31] we can observe that for any given method, including the ones that do not take disagreement into account during training, the CT F1 score is always higher than the F1 score. So one reason for choosing F1 or CT F1 is whether one finds her/himself in agreement with the argument by Dumitrache et al. (2018c) that evaluating a model under the assumption of a single correct answer underestimates its performance.

Another possible criterion for deciding between F1 and CT F1 is the observation that with CT F1, the gains of highly accurate (typically, gold-trained) models over less accurate models are reduced. This can be gathered from the results with five of the six datasets, the one exception being RTE. Consider, for example, the difference between the best gold/gold-plus method and the best crowd-only method on the POS dataset in Tables 8 and 9. The F1 difference between gold training and training using the KL soft-loss method was 11.14 points, while the CT F1 difference was 5.8. In other words, the difference we observed between the performance of the KL soft-loss method and the performance of the gold model was much reduced when confusion was factored in. One interpretation would be that if we prefer to report only one metric, CT F1 may give a picture of the differences among methods less affected by difficult items. (Alternatively, it might be possible to design an experiment to test which of the two metrics yields scores more in keeping with human intuition.)

The differences among the results according to soft evaluation metrics are more substantial. Both CE and JSD measure the distance between the probability distribution outputted

30. And of course, at the other extreme—which we didn't investigate in this study—of tasks where the labels are highly subjective, such as hate speech detection, it may be argued that using a hard metric makes little sense. For such tasks, it would appear that soft evaluation metrics such as those trialled here would be more appropriate.

31. We compare CT F1 with F1 rather than with accuracy as the two metrics differ only in the down-weighting of confusing items by CT F1 (see Section 4).

by a model and the target distribution, and the two measures are closely related, but apart from POS and IC-LABELME, these metrics yielded very different results. More research is required in order to understand which of the two methods more accurately reflects intuition, but given that cross-entropy is already widely used in practice, it would certainly be reasonable to interpret our results as not providing sufficient justification for choosing JSD over CE. The results for both these metrics were also substantially different from those obtained with the two entropy-based measures, but these also differed from each other to a large degree. Again, no conclusion can be reached at this point in time as to which of these metrics is more appropriate for the purposes of assessing how well models capture the uncertainty among human judgments. One conclusion is, however, clear: It is time to move beyond comparing models to a single ground truth (Basile et al., 2021).

9. Conclusion

With the growth in size, sophistication, and quality of annotated resources, and the increasing practice of employing several annotators to produce them, the idealization that a "gold" interpretation can be specified for every item in the dataset and used as the target during learning/evaluation, underlying much practice in supervised learning, is becoming less and less justifiable. (Also increasing is the evidence that training with noisy labels results in better performance of the obtained models on unseen data.) Abandoning this idealization requires the development of new paradigms both for training and for evaluating models. In this paper, we identified several AI tasks for which the gold-standard idealization has been shown not to hold and used them for an in-depth analysis of the by now extensive literature on learning from data possibly containing disagreements.

Our results suggest, first of all, that reaching a consensus on how to evaluate models if we abandon the gold-standard assumption is an essential prerequisite for this research, as the relative performance of the training methods under consideration is critically affected by the chosen evaluation. Our experiments do not allow us to reach a definitive conclusion in this matter, as no soft evaluation metric was found to be clearly more appropriate than any other. Until such consensus is reached, however, we found no reason not to simply use cross-entropy to compare the output of a system to a soft label.

Secondly, we observed a strong dataset effect. With datasets of a substantial size and providing large numbers of judgments for each item, annotated by high-quality coders, training directly from the soft labels achieved better results than training from aggregated labels, or even from gold labels, both when evaluating using hard evaluation and when using soft evaluation. When those conditions did not hold, leveraging gold labels generally achieved the best results in terms of hard evaluation, but leveraging soft labels in addition to gold labels generally achieved the best overall results, greatly improving performance when measured using soft metrics and leading to as good or better results than using gold labels only in terms of hard evaluation with datasets not satisfying the conditions discussed above. Among the methods not relying on a gold label, it was notable that aggregation generally resulted in worse performance than training directly from the soft label, particularly when using soft-loss or repeated-labelling methods.

In terms of recommendations for future directions, more research is clearly needed on the evaluation side. Other than that, our results suggest that the best way to achieve

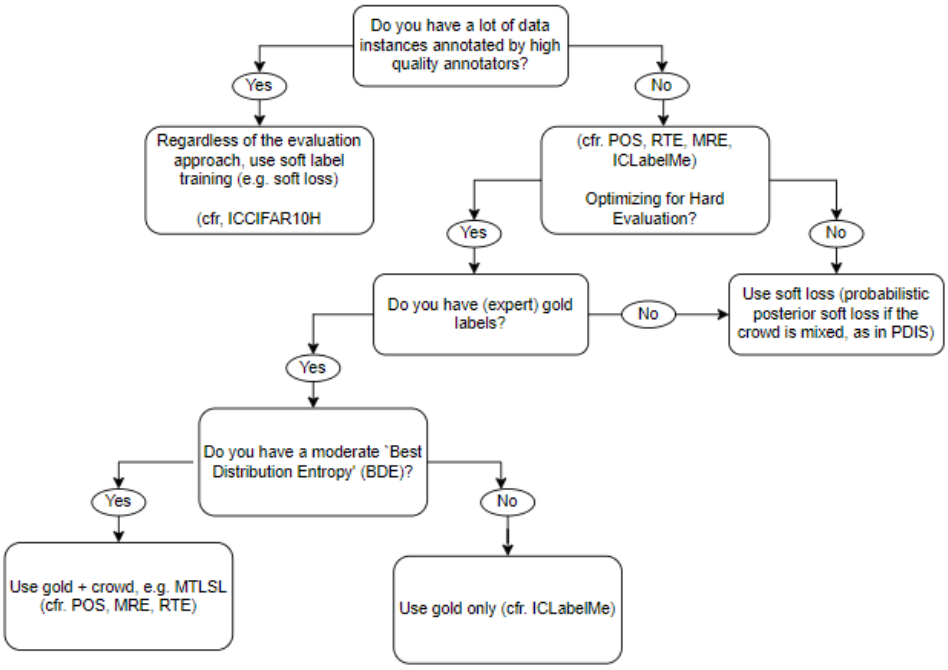


Figure 12: A guide to choosing a best-performing model given the characteristics of one's dataset and hard or soft evaluation

high-quality and empirically grounded datasets is to collect a substantial number of judgments from high-quality coders. Our recommendations for researchers working with existing datasets are summarized in the somewhat simplified decision tree in Figure 12. (We encourage researchers to read in detail the parts of this study discussing the datasets identified in the simplified decision tree that most resemble their own.)

Acknowledgments

We wish to thank our reviewers for two extremely detailed reviews that also contained many practical suggestions to improve the paper. Alexandra Uma, Silviu Paun, and Massimo Poesio were supported by the DALI project, ERC Advanced Grant 695662 to Massimo Poesio. Barbara Plank is in part supported by the Independent Research Fund Denmark (DFF) grant 9131-00019B and grant 9063-00077B.

References

Luis von Ahn and Laura A. Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51:58–67.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA - XVIIIth International Conference of the Italian Association for Artificial Intelligence*, Lecture Notes in Computer Science, page 588–603. Springer.

Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. 2016. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321.

Hector Martinez Alonso, Anders Johannsen, and Barbara Plank. 2016. Supersense tagging with inter-annotator disagreement. In *Proc. of LAW*, pages 43–48, Berlin. Association for Computational Linguistics.

Hector Martinez Alonso, Barbara Plank, Arne Skjaerholt, and Anders Sogaard. 2015. Learning to parse with IAA loss. In *Proc. of NAACL*, pages 1357–1361, Denver. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36:15–24.

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

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of preaggregation on the evaluation of highly subjective tasks. In *Proc. of the AIXIA Workshop*. Universitá di Torino.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Eyal Beigman and Beata Beigman-Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, Suntec, Singapore. Association for Computational Linguistics.

Beata Beigman-Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35:495–503.

Beata Beigman-Klebanov and Eyal Beigman. 2014. Difficult cases: from data to learning, and back. In *Proc. of the ACL*, pages 390–396.

Beata Beigman Klebanov, Eyal Beigman, and Danie Diermaier. 2008. Analyzing disagreements. In *Proceedings of the Coling 2008 workshop on Human Judgements in Computational Linguistics*, pages 2–7.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Bob Carpenter. 2008. Multilevel bayesian models of categorical data annotation. Available as `http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf`.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! born-again multi-task networks for natural language understanding. In *Proc. of EMNLP*, Hong Kong. ACL.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. Deliverable D16, The FRACAS Project.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *MLCW 2005*, pages 177–190. Springer.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, 51(1).

A. Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171—-4186, Minneapolis. ACL.

Anca Dumitrache. 2019. *Truth in Disagreement*. Ph.D. thesis, Free University Amsterdam.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018a. Crowdsourcing ground truth for medical relation extraction. *ACM Trans. Interact. Intell. Syst.*, 8(2).

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018b. Crowdsourcing semantic label propagation in relation classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proc. of NAACL*.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018c. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement (short paper). In *SAD/CrowdBias@HCOMP*.

Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Martinez-Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. 2018d. Empirical methodology for crowdsourcing ground truth. *ArXiv*, abs/1809.08888.

Paul Felt, Eric Ringger, Jordan Boyd-Graber, and Kevin Seppi. 2015. Making the most of crowdsourced document annotations: Confused supervised LDA. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 194–203, Beijing, China. Association for Computational Linguistics.

Michael Firman, Neill D. F. Campbell, Lourdes Agapito, and Gabriel J. Brostow. 2018. Diversenet: When one right answer is not enough. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5598–5607.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2004. *The Measurement of Interrater Agreement*, chapter 18. John Wiley & Sons, Ltd.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *ICML*.

Xavier Gastaldi. 2017. Shake-shake regularization of 3-branch residual networks. In *ICLR*.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. *ArXiv*, abs/1712.09482.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

Human Language Technologies, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Benjamin Graham. 2014. Fractional max-pooling. *ArXiv*, abs/1412.6071.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. 2017. Who said what: Modeling individual labelers improves classification. In *AAAI*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Geoffrey E. Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. ArXiv preprint arXiv:1503.02531.

Yufang Hou. 2016. Incremental fine-grained information status classification using attention-based lstms. In *COLING*.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Global inference for bridging anaphora resolution. In *HLT-NAACL*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, Baltimore, Maryland. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, page 57–60, USA. Association for Computational Linguistics.

Nancy Ide and James Pustejovsky, editors. 2017. *The Handbook of Linguistic Annotation*. Springer.

Oana Inel and Lora Aroyo. 2017. Harnessing diversity in crowds and machines for better NER performance. In *Proc. of ESWC*, pages 289–304. Springer.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *International Semantic Web Conference (ISWC)*, pages 486–504.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proc. of HLT-NAACL*, pages 556–562.

Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2015. Identifying and accounting for task-dependent bias in crowdsourcing. In *HCOMP*.

Kian Kenyon-Dean, Eisha Ahmed, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: it's complicated! In *Proc. of NAACL*, pages 1886–1895. ACL.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Olga Krasavina and Christian Chiarcos. 2007. The Potsdam coreference scheme. In *Proc. of the 1st Linguistic Annotation Workshop*, pages 156–163.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. *Games with a Purpose (GWAPs)*. Wiley.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

John P. Lalor, Hao Wu, and Hong Yu. 2018. Soft label memorization-generalization for natural language inference. In *Proc. of UAI UDL Workshop*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *NAACL-HLT*.

Yuan Li, Benjamin Rubinstein, and Trevor Cohn. 2019. Exploiting Worker Correlation for Label Aggregation in Crowdsourcing. In *International Conference on Machine Learning*, pages 3886–3895. ISSN: 1938-7228 Section: Machine Learning.

Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *ACL*.

Marie-Catherine de Marneffe and Christopher Potts. 2017. *Developing Linguistic Theories Using Annotated Corpora*, pages 411–438. Springer Netherlands, Dordrecht.

Pietro Michelucci, editor. 2013. *Handbook of Human Computation*. Springer.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, page 1003–1011, USA. Association for Computational Linguistics.

Volodymyr Mnih and Geoffrey Hinton. 2012. Learning to label aerial images from noisy data. In *Proc. of ICML*, Edinburgh, Scotland.

Pablo G. Moreno, Antonio Artés-Rodríguez, Yee Whye Teh, and Fernando Perez-Cruz. 2015. Bayesian nonparametric crowdsourcing. *J. Mach. Learn. Res.*, 16(1):1607–1627.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176, Portorož, Slovenia. European Language Resources Association (ELRA).

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proc. of LREC*.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2019. Confident learning: Estimating uncertainty in dataset labels. In `https://arxiv.org/abs/1911.00068`.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

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

Silviu Paun, Ron Artstein, and Massimo Poesio. 2021. *Statistical Methods for Annotation Analysis*. Morgan Claypool.

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

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *Computing Research Repository - CORR*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Massimo Poesio. 2020. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann, editors, *The Companion to Semantics*. Wiley.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proc. of LREC*, Marrakesh.

Massimo Poesio, Jon Chamberlain, and Udo Kruschwitz. 2017. Crowdsourcing. In N. Ide and J. Pustejovsky, editors, *The Handbook of Linguistic Annotation*, pages 277–295. Springer.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Massimo Poesio, Uwe Reyle, and Rosemary Stevenson. 2007. Justified sloppiness in anaphoric reference. In H. Bunt and R. Muskens, editors, *Computing Meaning*, volume 3, pages 11–34. Kluwer.

Massimo Poesio, Patrick Sturt, Ron Arstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes*, 42(2):157–175.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, page 1–27, USA. Association for Computational Linguistics.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Ellen F. Prince. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering – WISE 2013*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Marta Recasens, Ed Hovy, and M. Antonia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Proc. of the Workshop on Human Judgments in Computational Linguistics*, pages 8–16.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Arndt Riester, David Lorenz, and Nina Seeman. 2010. A recursive annotation scheme for referential information status. In *Proc. of LREC*, pages 717–722.

Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco Pereira. 2017. Learning supervised topic models for classification and regression from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proc. of AAAI*.

Bryan Russell, Antonio Torralba, Kevin Murphy, and William Freeman. 2008. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77.

Viktoriia Sharmanska, Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2194–2202.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Aashish Sheshadri and Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. In *HCOMP*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

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

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring ground truth from subjective labelling of venus images. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, page 1085–1092, Cambridge, MA, USA. MIT Press.

Rion Snow, Brendan O Connor, Daniel Jurafsky, Andrew Y Ng, Dolores Labs, and Capp St. 2008. Cheap and fast - but is it good? Evaluation non-expert annotiations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29, pages 1857–1865. Curran Associates, Inc.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806.

Manfred Stede. 2008. Disambiguating rhetorical structure. *Research in Language and Computation*, 6(3–4):311–332.

Georg Stemmer, Stefan Steidl, Elmar Nöth, Heinrich Niemann, and Anton Batliner. 2002. Comparison and combination of confidence measures. In *Text, Speech and Dialogue*, pages 181–188, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2019. Robust argument unit recognition and classification. *CoRR*, abs/1904.09688.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft-loss functions. In *Proc. of HCOMP*.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*, 26(1).

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6:333–353.

Chang Wang and James Fan. 2014. Medical relation extraction with manifold models. In *ACL*, volume 1, pages 828–838.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142.

Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc.

Hui Yang, Anne De Roeck, Vincenzo Gervasi, Alistair Willis, and Bashar Nuseibeh. 2011. Analysing anaphoric ambiguity in natural language requirements. *Requirements Engineering*, 16:163–189.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Jing Zhang, Xindong Wu, and Victor Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*.

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