Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement

Pia Sommerauer[*] **Antske Fokkens**[*†] **Piek Vossen**[*]

[*]Computational Lexicology and Terminology Lab, Vrije Universiteit Amstersterdam
De Boelelaan 1105, 1081HV Amsterdam
[†]Department of Mathematics and Computer Science, Eindhoven University of Technology
P.O.Box 513, MetaForum 5600 MB Eindhoven
{pia.sommerauer,antske.fokkens,p.t.j.m.vossen}@vu.nl

Abstract

Semantic annotation tasks contain ambiguity and vagueness and require varying degrees of world knowledge. Disagreement is an important indication of these phenomena. Most traditional evaluation methods, however, critically hinge upon the notion of inter-annotator agreement. While alternative frameworks have been proposed, they do not move beyond agreement as the most important indicator of quality. Critically, evaluations usually do not distinguish between instances in which agreement is expected and instances in which disagreement is not only valid but desired because it captures the linguistic and cognitive phenomena in the data. We attempt to overcome these limitations using the example of a dataset that provides semantic representations for diagnostic experiments on language models. Ambiguity, vagueness, and difficulty are not only highly relevant for this use-case, but also play an important role in other types of semantic annotation tasks. We establish an additional, agreement-independent quality metric based on answer-coherence and evaluate it in comparison to existing metrics. We compare against a gold standard and evaluate on expected disagreement. Despite generally low agreement, annotations follow expected behavior and have high accuracy when selected based on coherence. We show that combining different quality metrics enables a more comprehensive evaluation than relying exclusively on agreement.

1 Introduction

Would you say leopards are yellow? Most likely, some people would while others would not. Both interpretations are valid, as the interpretation depends on a person's boundaries for the properties 'yellow' and 'brown'. Selecting only one judgment would disregard the vagueness of the expression, a phenomenon at the heart of lexical semantics. At the same time, most people would probably agree that wine can be red without having to think about it. A high number of semantic annotation tasks is characterized by unclear, difficult, ambiguous and vague examples. Annotation, in particular when distributed among a crowd, has the potential of capturing different interpretations, conceptualizations and perspectives and can thus provide highly relevant semantic information. Existing evaluation and label extraction methods, however, still heavily rely on agreement between annotators, which implies a single correct interpretation. Finished datasets rarely provide indications about difficulty and ambiguity on the level of annotated units.

The explanatory power of NLP experiments that aim to evaluate or analyze models depends on the informativeness of the data. This is particularly relevant for experiments which specifically aim to understand models better, such as the tradition of diagnostic experiments (Belinkov and Glass, 2019).Traditional error analyses could also benefit substantially from test sets which contain information about phenomena with a likely impact on model performance. Furthermore, knowing whether model-errors are similar to human disagreements can yield important insights about models. For instance, an analysis of natural language inference models shows that classifiers do not necessarily capture the same type of

Proceedings of the 28th International Conference on Computational Linguistics, pages 4798–4809
Barcelona, Spain (Online), December 8-13, 2020

ambiguity and uncertainty as reflected in the annotations (Pavlick and Kwiatkowski, 2019). Error analysis often require manual annotation and tend to focus on small and not representative subsections of test sets (Wu et al., 2019). We argue that the behavior of human annotators can provide rich information which should be exploited, rather than reduced to single labels. Information about (dis)agreement is a by-product of the original annotation effort and thus comes for free. It can form the basis of an error analysis or, in the case of our data, should be used to draw informative conclusions from diagnostic experiments. Such experiments crucially depend on the quality and informativeness of the underlying data (Hupkes et al., 2018).

In this paper, we present an approach to crowd-annotation for a diagnostic dataset which attempts to tackle these limitations. The dataset is meant to test which semantic properties are captured by distributional word representations. The task is designed to trigger fine-grained semantic judgements of potentially ambiguous examples. The behavior of ambiguous words in distributional semantic models is not well understood and thus particularly interesting (Sommerauer and Fokkens, 2018; Yaghoobzadeh et al., 2019; Del Tredici and Bel, 2015). We investigate to what extent existing and new quality metrics indicate annotation accuracy on the one hand and ambiguity and difficulty of annotation units on the other hand. We evaluate our task from three perspectives: (1) comparison against an expert-annotated gold standard, (2) a task-specific coherence metric independent of agreement and (3) evaluation in terms of inter-annotator agreement metrics compared to predefined expectations about agreement and disagreement. In particular, we aim to investigate (1) how we can exploit the strengths and weaknesses of various suggested metrics to select and aggregate labels provided by the crowd, (2) to what degree disagreement among workers occurs in cases where it is expected and legitimate and (3) which metrics are suitable for detecting annotation units with legitimate and informative disagreement.[1]

Disagreement has been shown to indicate ambiguous cases when measured with the CrowdTruth framework (Aroyo and Welty, 2014; Dumitrache et al., 2018). However, we are not aware of work which compares different (dis)agreement and difficulty metrics. To the best of our knowledge, there is no study which tests how well different metrics can be used to identify ambiguous annotation units in a set of units annotated in terms of expected and legitimate disagreement. We show that the metrics we use give complementary insights and can be used to filter and aggregate labels in a way that produces high-quality annotations. Despite a relatively low inter-annotator-agreement, we show that worker behavior follows our expectations about agreement and disagreement and that high-quality labels can be extracted from the annotations, in particular for cases where we expect worker agreement.

The remainder of this paper is structured as follows: After reviewing related work (Section 2), we introduce the use-case of a diagnostic dataset (Section 3) and describe the annotation task (Section 4). We present our expert-annotated gold standard in Section 5 and different quality metrics in Section 6. The results of our experiments are described in Section 7, followed by a discussion and conclusion.

2 Related work

Recent annotation studies recognize that ambiguity, vagueness and varying degrees of difficulty are inherent to semantic phenomena (Dumitrache et al., 2019; Aroyo and Welty, 2015; Erk et al., 2003; Kairam and Heer, 2016; Poesio et al., 2019; Pavlick and Kwiatkowski, 2019). Pavlick and Kwiatkowski (2019) demonstrate that the fundamental task of Natural Language Inferencing contains large proportions of instances with multiple valid interpretations and argue that this phenomenon is central to the task rather than an aspect which can be disregarded. Herbelot and Vecchi (2016) show that even experts disagree on a difficult semantic annotation task and that interpretations are likely to vary due to differences in conceptualizations, which are in themselves justified and cannot simply be disregarded as ‘mistakes’.

Information about ambiguity and difficulty is crucial for a number of areas in NLP. While traditional experiments tend to accept the level of informativeness provided by standardly used corpora, the rather recent trend of model analysis (Belinkov and Glass, 2019) particularly highlights the necessity of informative data. Such approaches test to what extent uninterpretable, machine or deep learning-based

[1] The crowd and expert annotations are available at this repository: `https://github.com/cltl/SPT_crowd_data_analysis`

models capture certain aspects of information. They are highly experimental and crucially depend on the quality and informativeness of the diagnostic dataset (Hupkes et al., 2018; Sommerauer and Fokkens, 2018). In some cases, establishing a clean dataset is less problematic, for example when considering relatively clear and uncontested aspects of linguistic analysis (e.g. part-of-speech information (Saphra and Lopez, 2018)) or when creating entirely artificial data (Hupkes et al., 2018). Approaches which aim to capture information about semantics (such as embedding analysis (Sommerauer and Fokkens, 2018; Yaghoobzadeh et al., 2019)), however, are much more complex as ambiguity, vagueness and differences in required knowledge are by no means marginal phenomena and cannot simply be disregarded. Furthermore, the role of ambiguity in the behavior of word embeddings is not fully understood yet (Del Tredici and Bel, 2015; Yaghoobzadeh et al., 2019). We have designed an annotation task to analyze how different aspects of word meaning are represented in distributional representations (Sommerauer et al., 2019). In this paper, we investigate how we can measure the quality of the annotations and capture valid disagreement, which is crucial information for the diagnostic experiments we want to conduct (Sommerauer, 2020). The task is similar to that of Herbelot and Vecchi (2016), but uses basic yes-no questions so that it is suitable for crowd-annotations. It includes more fine-grained semantic judgments and intentionally ambiguous words. We can thus expect even more disagreement than already observed in Herbelot and Vecchi (2016).

Despite the central nature of phenomena triggering disagreement in annotation tasks, we are not aware of evaluation methods that do not mainly rely on agreement. Traditionally, annotations by a few annotators who worked on the same units are evaluated in terms of Kappa scores (usually Cohens's kappa) and tasks with varying workers annotating the same units (usually crowd tasks) in terms of Krippendorff's alpha (Artstein and Poesio, 2008). The CrowdTruth framework suggested in Aroyo and Welty (2014) and Aroyo and Welty (2015) offers a more fine-grained view by distinguishing the levels of workers, units and labels, rather than reducing the entire task to a single score. The goal is to distinguish meaningful disagreements (i.e. agreements by reliable annotators) from noise (i.e. disagreement or agreement by generally unreliable annotators). The framework provides scores for workers, annotation units (clear units receive a high score, units triggering disagreement between reliable annotators a low score), labels and associations between units and labels. The scores can be used to aggregate labels and for identifying unclear annotation units, as for instance shown in Dumitrache et al. (2015) and Dumitrache et al. (2019). Other approaches attempt to discover disagreeing but valid interpretations in annotations based on clustering (Kairam and Heer, 2016) and Gaussian modeling (Pavlick and Kwiatkowski, 2019). While these approaches provide valuable insights, we focus on transparent and simple methods for quality assessment which do not require a large volume of data.

3 Use case: a dataset for diagnostic experiments

In this section, we present our use-case, which requires particularly high-quality data. Annotations should be provided by a crowd, rather than experts, as we are interested people's general perception rather than expert judgments. Though we focus on a task with these specific characteristics, we believe that the general approach presented in this paper can also yield important insights in other, perhaps more traditional annotation scenarios.

Experiments in the tradition of model analysis require informative and high-quality data, as they aim to discover general tendencies about what kind of information models can capture. Sommerauer et al. (2019) propose a dataset for diagnostic experiments on word embeddings using property-concept pairs annotated with fine-grained semantic judgments. The dataset is meant to test whether a semantic property (e.g. 'flying') is encoded by embedding representations or not. This can be investigated by testing whether positive (e.g. 'seagull', 'airplane', 'bee') and negative candidate concepts (e.g. 'penguin', 'train', 'ant') can be distinguished purely based on their embedding. The examples should not only be used to test whether a specific semantic property is encoded in embeddings, but, beyond this, help to uncover underlying factors determining whether a property is reflected in a distributional representation of a concept or not. Therefore, the concept-property pairs should be annotated with semantic relations reflecting various linguistic factors. Each concept-property pair can be connected by one or more of a total of ten relations (for instance expressing types of typicality or whether there can be variation in instances of concepts). The semantic relations can be grouped with respect to the subset of concept-instances a property applies to (most to all, some, or few to no instances of a concept). This enables diagnostic experiments with positive and negative examples. We encountered the problem of annotation evaluation given expected disagreement while compiling this dataset as it contains a high number of ambiguous instances and instances of varying degrees of difficulty, for which disagreement can be valid and meaningful. In an ideal scenario, our analysis of annotations can (1) provide an overall indication of annotation quality which does not purely rely on agreement and (2) distinguish different types of disagreement. At the most coarse-grained level, it should distinguish justified disagreement from noise (caused by mistakes or spammers).

4 Annotation task

The goal of the annotation task is to annotate property-concept pairs with relations. To make the task simple and suitable for a crowd of untrained workers, we turned it into a binary-decision task. This means that a single annotation unit consists of a property-concept-relation triple. This results in ten annotation units per concept-property pair. As the relations have rather abstract names, we translate them to natural language statements describing a property and a concept. The following statement is an example of a description expressing the property-concept-relation triple **black**-*rhino*-`variability_limited`: *You can find (a/an) rhino which is black. Black is one of a few possible colors (a/an) rhino usually has. There is only a limited range of possible colors.* Participants are asked to indicate whether they agree with a given statement about a property and a concept. More example statements are listed in Table 1.[2]

To avoid triggering random answers, we encourage participants to look up words they do not know. Each statement is introduced by a short instruction sentence and an example of the same relation and property-type which would most likely trigger the response ‘agree’ and which would trigger ‘disagree’.

relation	example
typical_of_concept	“**Spicy**” is one for the first things which come to mind when I hear “chili pepper’ because **spicy** is one of the typical tastes of (a/an) *chili pepper*’.
typical_of_property	“*Feather*” is one of the first things which come to mind when I hear “**light**’ because (a/an) *feather* is a typical example of things which are **light**’.
affording_activity	I know that having (an/an) **blade** is necessary for many things (a/an) *razor* does or is used for.
variability_open	You can find (a/an) *t-shirt* which is **white**. White is one of many possible colors (a/an) *t-shirt* usually has. The range of colors is almost unlimited.
rare	I think (a/an) *wine glass* can is **made of plastic**, but this is rare or uncommon.
impossible	I think it is impossible for (a/an) *corpse* to be **alive**.

Table 1: Examples of statements expressing semantic relations.

We used the freely available Lingoturk software (Pusse et al., 2016) to set up an annotation environment and distributed the task via the recruitment platform Prolific.[3] Peer et al. (2017) show that the annotation quality of annotators recruited via Prolific is higher than for Amazon Mechanical Turk workers. The platform encourages fair payment and asks researchers to pay participants based on the time they estimate for a task rather than per annotated item.

We split the dataset into batches of around 70 descriptions. A worker who is proficient in English would need about 10 minutes per batch. While some statements may be difficult to judge and therefore take more time, most are expected to be rather intuitive and easy to answer. Annotators were paid based on the UK minimum wage. Each unit was annotated by 10 workers. To enable regular quality checks, we always include the full range of descriptions associated with a property-concept pair in the same batch. This enables us to check whether answers contradict each other. It has the disadvantage that the diversity of property-concept pairs in a batch is low.

[2] The entire set of input data can be found at this repository: `https://github.com/cltl/SPT_annotation`

[3] `https://www.prolific.co/`

We monitored the quality of the annotations during the annotation process and used inter-mediate worker evaluations to 'recruit' good annotators. Rather than rejecting low-quality submissions, we developed a 'whitelist' approach. Prolific enables researchers to distribute studies exclusively among a pre-selected group of workers. We test whether workers contradict themselves in their answers (explained in more detail below), for instance by judging a property as typical of a concept and at the same time stating that it is unusual of the concept. As we do not know how much legitimate disagreement could be expected in a single batch, we decide to rely on an agreement-independent metric rather than inter-annotator agreement.

5 A gold standard for accuracy and expected agreement

We establish a gold standard to evaluate (1) the accuracy of annotations extracted based on different quality metrics and (2) the ability of different metrics to identify justified and potentially meaningful disagreement. The authors of the paper annotated a subset of already annotated units. The units for expert annotation were selected from units with high, medium and low agreement. Agreement was established by calculating Krippendorff's alpha on the level of concept-property pairs (each pair has up to ten units).[4] This resulted in a set of 154 units (containing 19 property-concept pairs and 11 different properties). The inter-annotator agreement before discussion was 0.51 and 0.72 after discussion (averaged pairwise Cohen's kappa). We count all units in which agreement between experts could not be reached as units with expected disagreement. These units (23 in total) are excluded from the gold standard for label accuracy, as there are no incorrect answers in these cases.

We also indicated whether we expected the crowd to disagree for legitimate reasons. Examples of such disagreements are shown in Table 2. We identify different reasons for expected disagreement, such as vagueness in the property, ambiguity in either the concept or the property, odd property-concept combinations, etc.). We used these categories to facilitate the expert annotation process. While they served as a helpful tool for annotation and discussion, the inter-annotator agreement with respect to the disagreement categories remained low. It has to be considered that in most cases, various categories interact. When discussing annotations, we could frequently reach agreement about the subset of disagreement categories involved in an annotation unit, but disagreed about where the emphasis should be placed. In our current analysis, we simplify and distinguish the following three categories: agreement, possible disagreement and almost certain disagreement. Agreement was chosen for cases where all annotators expected agreement, possible disagreement for mixed cases and disagreement for cases where all annotators indicated they expected disagreement. We argue that taking these unions is most sensible, as multiple perspectives are necessary to discover possible reasons for disagreement. In total, we expect agreement for 49 units, possible disagreement for 48 and almost certain disagreement for 57 units. For 23 of the 57 units, a gold label could not be reached in expert discussion.

6 Quality metrics

We experiment with three types of quality metrics: We consider traditional inter-annotator agreement, quality scores in the CrowdTruth framework and our own, task-specific coherence metric. The metrics assess different aspects of the annotated dataset, as explained below.

6.1 Traditional inter-annotator-agreement

Traditionally, annotation tasks are assessed in terms of inter-annotator agreement (Artstein and Poesio, 2008). Crucially, inter-annotator agreement metrics should go beyond simple ratios and account for the possibility of agreement by chance. Widely used scores which do this are Cohen's Kappa (suitable for pair-wise assessment of annotators) and Krippendorff's alpha (suitable for a large number of annotators who are not consistent across the set). Both scores range between -1 and 1. Artstein and Poesio (2008) argue that Computational Linguistics tasks should require an agreement of 0.8 (while agreement above 0.67 is generally considered acceptable for some tasks). Such a strict threshold would not do justice to our task, which is characterized by expected ambiguity and disagreement. Traditionally, these metrics are

[4] For some pairs, some relations were excluded based on existing annotations from Herbelot and Vecchi (2016).

Vague property	The property is vague. *Usually, (a/an) leopard is not yellow, but there could be a highly unusual situation in which (a/an) leopard is yellow.*
Ambiguous property	The property is ambiguous and not disambiguated in the context of the concept and description. *You can find (a/an) chutney which is hot. (A/an) chutney is usually either hot, a bit more or less hot or the opposite of hot.*
Ambiguous concept	The concept is ambiguous and not disambiguated in the context of the property and description. *I know that (a/an) trumpeter can fly/be used for flying as most or all other things similar to (a/an) trumpeter fly.*
Odd pair	The combination of the property and concept is strange and confusing. This is always the case, regardless of the description. *You can find (a/an) recliner which is square. Square is one of a few possible shapes (a/an) recliner usually has. There is only a limited range of possible shapes.*
Odd triple	The combination of the property, concept and description is strange and confusing. *I know that being yellow is necessary for many things (a/an) buttercup does or is used for.*
Differences in conceptualization	The description asks the participant to place the concept in a conceptual system. The answer depends on the conceptual system of the participant. *I know that (a/an) arrow can fly/be used for flying as most or all other things similar to (a/an) arrow fly.*
Specialized knowledge	Answering this correctly requires specialized knowledge. It is likely that not all workers are aware of this. *I think (a/an) carrot can be red, but this is rare or uncommon.*
Imagination	This depends on how creative and imaginative a participant is. This type of disagreement only matters for confusions between negative relations (e.g. rare, unusual, impossible). *I think there is a shovel which can roll/be used for rolling, but this is rare or uncommon.*

Table 2: Expected reasons for worker disagreement.

used to give indications about the quality of the full set. In contrast, we use them directly to investigate whether expected disagreement indeed leads to lower alpha scores.

6.2 CrowdTruth metrics

The CrowdTruth framework was specifically designed to account for ambiguity and different levels of difficulty in a crowd-annotation setting. Beyond accounting for variation in the data, it also considers that crowd workers may have different abilities and that labels used in the annotation process can vary with respect to clarity. Rather than using a single aggregated score, the framework proposes metrics for workers, annotation units, labels and association strength between units and labels. Each task-component (workers, units and labels) is represented by a vector. The scores are calculated in terms of cosine similarities (expressing agreement) and weighted. For example an annotation unit on which most workers disagree receives a lower weight, just like a worker who frequently disagrees with other workers. Each score can take a value between 0 and 1. Dumitrache et al. (2019) show how the individual scores can be used for label identification and the identification of ambiguous units. The unit-quality-score (uqs) measures the weighted worker agreement on a particular unit and can be used to identify unclear or difficult units. The unit-annotation score (uas) measures the weighted agreement on a particular label for a unit. This indicates which label should be selected based on the analysis. Finally, we experiment with the worker quality score (wqs) for filtering low-quality workers.[5]

6.3 Task-specific metric: contradiction ratio

We define a metric specific to our task which assesses the coherence of worker judgments independent of agreement. We assume that reliable workers should not contradict themselves in the judgments of units associated with a single property-concept pair. For example, stating that a **fly** is typical of *penguin* and that it is impossible that *penguins* **fly** would count as a contradictory annotation. The semantic relations associated with a single pair can be divided into relations expressing that a property applies to all or most concept-instances, some concept-instances or few to no concept-instances. Contradictory annotations are annotations which state that relations in the `most/all`-category and the `few-none` category are true. We calculate a contradiction rate by dividing all observed contradictions by all possible contradictions for a property-concept pair. This can be done for the annotations of an individual worker or all annotations for a pair. The contradiction rate for the worker can be seen as an indication of worker quality (the lower

[5] We use the scores as they are defined in the appendix of Dumitrache (2019).

the better). The contradiction ratio on the level of a pair can be seen as an indication of the difficulty of a property-concept combination (the higher the more difficult).

7 Results

In this section, we present the results of our analysis. Section 7.1 presents a general overview and statistics about the collected annotations. In Section 7.2, we show the results of our evaluation against the gold standard in terms of label accuracy, followed by our evaluation with respect to expected agreement and disagreement (Section 7.3). Finally, we test how well different quality metrics are able to identify units with legitimate disagreement (Section 7.4).

7.1 Overview

Table 3 shows the overview of the current state of our dataset. The table shows statistics for three intermediate versions and the total dataset. In total, we have collected almost 200 000 annotations for almost 2000 property-concept pairs covering 13 different semantic properties with on average 150 associated concepts each. On average, each worker annotated about 183 units, which is more than two batches (of 70 questions each). The total inter-annotator agreement (measured by Krippendorff's alpha) is 0.31. If relations are merged into `most-all`, `some` and `few-none`, inter-annotator agreement rises to 0.37. If just the relations in the category `few-none` are merged, the alpha score is 0.33. We improved the formulation based on the outcome of our first runs. The first two intermediate versions have lower agreement scores than the third version as a result. The number of contradictions also declines (partly due our whitelist approach).

data	total	version1	version2	version3
n_annotations	195619.00	20971.00	41447.00	133201.00
n_properties	13.00	3.00	3.00	10.00
n_pairs	1935.00	425.00	426.00	1501.00
n_workers	1068.00	285.00	547.00	455.00
n_units	17907.00	4105.00	4094.00	13212.00
n_workers_per_unit	10.92	5.11	10.12	10.08
n_concepts_per_property_mean	148.85	141.67	142.00	150.10
annotations_per_worker_mean	183.16	73.58	75.77	292.75
iaa_label	0.31	0.23	0.21	0.36
iaa_collapse_neg	0.33	0.24	0.26	0.38
iaa_merged	0.37	0.22	0.24	0.43
contradiction_rate_mean	0.04	0.09	0.08	0.02

Table 3: Dataset overview.

7.2 Label accuracy

In this section, we present the results of the evaluation with respect to the correctness of extracted and aggregated crowd annotations compared to expert annotations. We experiment with different filtering and aggregation methods using the metrics described in Section 6.

Filtering. We filter based on worker-quality metrics (wqs and contradiction rate). Both scores require thresholds. We experiment with different thresholds calculated in terms of n standard deviations +/- mean calculated over the entire dataset, a batch or a single property-concept pair. Annotations made by workers with scores outside of the threshold are removed. We vary n between 0.5 and 2 (in steps of 0.5).

Aggregation methods. We use three different strategies for aggregation: Majority vote (a relation applies if $>$50% of workers select 'agree'), top vote (only the relation or, in case of a tie, the relations with the most 'agree' votes per pair) and varying unit-annotation score (uas) thresholds (between 0.5 and 1 in steps of 0.05). The top vote has the limitation that it usually only selects a single relation per pair as true, which disregards the nature of the task.

Results. Table 4 shows the weighted f1-scores for the full set of gold annotations. In total, the set includes 131 units with a gold label (21 positive and 110 negative). The combination of filtering and aggregation methods and their thresholds results in a high number of configurations. We only report the top three results and the best result for each filtering-aggregation possibility.[6] All filtering methods result in full coverage for the entire gold standard set. The results show that a majority vote on labels filtered by contradiction rates yield the highest performance (f1 between 0.88 and 0.85). In contrast, a simple majority vote achieves an f1-score of 0.78. The best CrowdTruth method (unit-annotation-score) achieves an f1-score of 0.84, which is comparable to removing all annotations containing contradictions. Using the worker-quality-score to exclude annotations does not improve results compared to a simple majority vote on unfiltered data. As can be expected, majority vote performs better than top vote.

When considering the f1-scores in comparison to the inter-annotator agreement, it can be seen that high performance does not necessarily depend on high agreement.

filtering	filtering_unit	n_stdv	aggregation	f1	p	r	alpha
contradictions	batch	0.5	majority_vote	0.88	0.92	0.86	0.21
contradictions	total	1	majority_vote	0.86	0.89	0.85	0.21
no_contradictions	-	-	majority_vote	0.85	0.87	0.85	0.34
-	-	-	uas-0.65	0.84	0.84	0.84	0.19
ct_wqs	batch	0.5	majority_vote	0.81	0.86	0.79	0.19
-	-	-	majority_vote	0.81	0.86	0.79	0.19
contradictions	pair	0.5	top_vote	0.79	0.82	0.78	0.21
no_contradictions	-	-	top_vote	0.78	0.82	0.76	0.34
ct_wqs	pair	1.5	top_vote	0.77	0.79	0.76	0.19
-	-	-	top_vote	0.77	0.79	0.76	0.19

Table 4: Evaluation of filtered and aggregated labels against expert annotations using precision, recall and a weighted f1-score. IAA is indicated by Krippendorff's alpha.

7.3 Expected crowd behavior

We compare the performance and inter-annotator agreement against expected agreement and disagreement. If the annotations reflect the data accurately, clear units should achieve a higher agreement than unclear, potentially ambiguous or difficult cases. Similarly, accuracy for clear cases should be high.

Table 5 lists the results for units in the gold set with expected agreement and the gold set with expected disagreement. In total, there are 49 units with expected agreement and 82 with expected disagreement (we merged possible and certain disagreement). For reasons of space, we only show the top three configurations, the top-configurations on the full set and some baseline configurations (majority vote on full, unfiltered set and excluding contradictory annotations). The inter-annotator agreement confirms the expected behavior (0.23 on the full set with expected agreement and 0.16 on the full set with expected disagreement). The results indicate that the contradiction-based filtering methods achieve high performance on both the set with expected agreement and expected disagreement, with only a slight advantage on the expected agreement set. The CrowdTruth unit-annotation-score (uas) methods perform highly on the set with expected agreements and drop on the set with expected disagreements (0.91 vs 0.79). We thus conclude that the contradiction-based methods provide a robust outcome and uas (CrowdTruth) can reflect differences in difficulty between sets.

A limitation of this comparison is that the two sets differ in size and balance of labels, which should be improved in an ideal set-up. The difference in inter-annotator agreement seems to be large enough to confirm that the workers behaved as expected. The results also indicate that robust labels can be extracted from a difficult set relying on contradiction-filtering.

7.4 Identifying units with valid disagreement

In this section, we investigate whether we can identify valid disagreement and distinguish it from noise. We evaluate how well unit-based quality metrics can distinguish units with expected disagreement from units with expected agreement. For this aspect of the evaluation, we use a stricter standard for identifying

[6] The full set of configurations and their results is included in the Github repository `https://github.com/cltl/SPT_crowd_data_analysis`.

expectation	aggregation	filtering	filtering_unit	n_stdv	f1	p	r	alpha
agree	majority_vote	contradictions	pair	0.5	0.91	0.94	0.90	0.28
agree	uas-0.65	-	-	-	0.91	0.92	0.90	0.23
agree	uas-0.7	-	-	-	0.90	0.91	0.90	0.23
agree	majority_vote	contradictions	total	1	0.89	0.94	0.88	0.28
disagree	majority_vote	contradictions	batch	0.5	0.89	0.91	0.88	0.16
agree	majority_vote	contradictions	batch	0.5	0.86	0.93	0.84	0.28
agree	majority_vote	no_contradictions	-	-	0.86	0.93	0.84	0.32
disagree	majority_vote	contradictions	batch	1	0.85	0.89	0.84	0.15
disagree	majority_vote	contradictions	batch	1.5	0.84	0.88	0.83	0.16
disagree	majority_vote	contradictions	total	1	0.84	0.86	0.83	0.17
disagree	majority_vote	no_contradictions	-	-	0.83	0.84	0.83	0.31
agree	majority_vote	-	-	-	0.83	0.92	0.80	0.23
disagree	majority_vote	-	-	-	0.81	0.83	0.79	0.16
disagree	uas-0.65	-	-	-	0.79	0.79	0.80	0.16

Table 5: Evaluation of aggregated labels against expert annotations for expected agreement and disagreement in terms of precision, recall and a weighted f1-score. IAA is indicated by Krippendorff's alpha.

expected disagreement in the expert annotations: We only use units which each of the expert annotators indicated as triggering disagreement and units with expected agreement. This leaves us with 49 units with expected agreement (as above) and 41 units with expected (and legitimate) disagreement.

We experiment with the unit quality score (uqs), proportional agreement (prop) and the contradiction rate. The latter two can be applied to the raw and filtered dataset (we use the best performing filtering method). For each metric, we calculate a threshold by establishing the mean over all units and test performance using mean +/- n * standard deviation. The best scores for each metric are reported in Table 6. We report the accuracy for identifying valid disagreement in comparison to the micro f1-score. The best result is achieved by using simple, proportional agreement on the dataset where contradictory annotations were removed. The contradiction rate on its own is not suitable for identifying difficult instances.

metric	n_sd +/-mean	accuracy (disagreement)	micro f1
uqs	0	0.68	0.50
prop	0	0.71	0.48
prop_filtered	0.5	**0.68**	**0.59**
contradictions	1	0.32	0.41
contradictions_filtered	1	0.32	0.41

Table 6: Accuracy of different metrics in identifying units with certain disagreement. Each metric requires a threshold, which we calculate based on mean +/- n standard deviations.

8 Discussion

In this paper, we have attempted to fill the gap between heavy emphasis of inter-annotator agreement on the one hand and justified disagreement on the other hand. Semantic annotation tasks have been acknowledged to contain ambiguous, difficult, vague and possibly confusing examples which are likely to trigger disagreement. While some approaches may still see these cases as marginal, we argue that they are a vital part of many linguistic phenomena and can yield important insights. In this paper, we have illustrated an approach for a dataset used in model analysis experiments. The tradition of model analysis methods places strong emphasis on the quality and soundness of datasets and the phenomena indicated by disagreement are particularly relevant for our task. However, we argue that datasets used in other experiments should be held to similarly high standards. The explanatory power of evaluation datasets for semantic tasks in general could be improved by explicitly containing information about disagreement.

We have shown that, for our particular use-case, the agreement-based metrics should not be used as the sole indicator of quality. Our results show that a task-inherent coherence check can yield important insights and serve as a valuable basis for discarding noisy annotations. While we have only shown this for our use-case, we believe that the principle can be applied to other annotation tasks as well. For example, we could imagine that the principle of logical coherence checks can be applied to a semantic role-labeling task. Predicates with contradictory semantic roles (based on the idea of selectional preferences) can be used as an indication of either noisy annotations or ambiguous annotation units. Even tasks that are particularly drawn to high disagreement, such as tasks in the domain of sentiment annotation, could benefit from such checks. In hate speech identification, it could be considered to check if (1) the same annotator uses opposing labels for very similar instances and (2) annotators completely contradict one another on the same instances (rather than just disagreeing about the boundaries of categories (such as 'positive' and 'neutral'). We do not intend to disregard the complex nature of such a task; other contextual factors, such as the background of the annotators, can also trigger contradictions. Taking these factors into account can yield further useful insights when interpreting (differences in) annotations. We believe that considering the interaction between these factors and logical checks can provide a valuable tool for analyzing and processing annotations.

While the approach presented here can be taken as a first step, there are still a number of limitations and remaining challenges. Most importantly, it would be highly valuable if the existing metrics could be combined in such a way that we could use them for the identification of different types of disagreements. For instance, it is relevant whether workers disagree because some have more specialized knowledge than others or because the annotation unit under consideration is indeed ambiguous. It could be considered to combine different metrics in such a way that they can distinguish between disagreement due to noise, disagreement because of differences in knowledge and disagreement due to real ambiguity. A possible way to achieve this could be to use the different metrics as features in a machine learning system. This research direction would require a larger volume of expert annotated gold data.

9 Conclusion

Despite the limitations discussed above, we draw the following conclusions: (1) Absolute thresholds for inter-annotator agreement and aggregated scores over all annotations disregard the nature of a difficult semantic task with ambiguous and vague instances. Rather, evaluations should focus on whether agreement can be found in cases where agreement can be expected. Our evaluation against expected agreement and disagreement shows that worker-behavior is in line with our expectations despite overall low inter-annotator agreement. (2) The results indicate that a simple, coherence-based task-specific worker-quality check yields accurate labels, even on datasets with low inter-annotator agreement. The advantage of this check is that it does not require high volumes of data to be accurate, but can be used with only a handful of annotated units. We expect that similar checks can also be established for other tasks. Such checks can be a cheap but high-impact approach, as they can be designed in such a way that they adhere to what is important in a particular task. In our case, good workers should understand questions and not contradict themselves. This is more important than that they agree with other workers. (3) High inter-annotator agreement is *not necessarily* a requirement for obtaining high-quality labels. Our evaluation shows that the highest f1-score on the expert-annotated gold standard was achieved by a filtering and aggregation method which does *not* result in the highest alpha score on the remaining labels. (4) While our approach to identifying legitimate disagreements is preliminary, we observe that a simple, proportional agreement metric on a dataset filtered for contradictory answers yields the best results. This research provides the groundwork for establishing the exact status of individual annotation units and thereby establish whether the information and quality is sufficient for experiments with computational linguistic models. As the size of our dataset increases, we plan to take the next steps towards such a fine-grained assessment.

Acknowledgements

This research is funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (NWO) PGW.17.041 awarded to Pia Sommerauer. We would like to thank Chantal van Son for her collaboration on the annotation set-up and anonymous reviewers for feedback that helped improve this paper. All remaining errors are our own.

References

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Human Computation*, 1(1).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

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

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Marco Del Tredici and Núria Bel. 2015. A word-embedding-based sense index for regular polysemy representation. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 70–78.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015. Crowdtruth measures for language ambiguity. In *Proc. of LD4IE Workshop, ISWC*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. 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 2164–2170.

A Dumitrache. 2019. Truth in disagreement: Crowdsourcing labeled data for natural language processing.

Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2003. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 537–544.

Aurélie Herbelot and Eva Maria Vecchi. 2016. Many speakers, many worlds. *LiLT (Linguistic Issues in Language Technology)*, 13.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648.

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

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

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.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *Proceedings of NAACL-HLT*.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2019. Towards interpretable, data-derived distributional semantic representations for reasoning: A dataset of properties and concepts. In *Wordnet Conference*, page 85.

Pia Sommerauer. 2020. Why is penguin more similar to polar bear than to sea gull? analyzing conceptual knowledge in distributional models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 134–142, Online, July. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763.

Yadollah Yaghoobzadeh, Katharina Kann, Timothy J Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753.