Empirical Analysis of Aggregation Methods for Collective Annotation

Ciyang Qing, **Ulle Endriss**, **Raquel Fernández** and **Justin Kruger**
Institute for Logic, Language and Computation
University of Amsterdam
{qciyang | justin.g.kruger}@gmail.com
{ulle.endriss | raquel.fernandez}@uva.nl

Abstract

We investigate methods for aggregating the judgements of multiple individuals in a linguistic annotation task into a collective judgement. We define several aggregators that take the reliability of annotators into account and thus go beyond the commonly used majority vote, and we empirically analyse their performance on new datasets of crowdsourced data.

1 Introduction

Human annotation of linguistic resources has become indispensable in computational linguistics, especially with regards to semantic and pragmatic information, which is yet beyond the reach of robust automatic labelling. Most annotation campaigns involve a small group of trained annotators who may not always agree on their judgements. The reliability of the annotation is typically assessed by quantifying the level of inter-annotator agreement, while the final annotation to be released is consensuated amongst experts. In recent years, however, crowdsourcing methods such Amazon's Mechanical Turk (AMT) have shaken up this scenario by making it possible to rapidly recruit large numbers of untrainned annotators at a low cost. This offers great opportunities—in particular, if we consider that the community of speakers is the highest authority regarding linguistic knowledge—but also creates several challenges: amongst others, how to obtain good quality annotations from untrainned and unmonitored individuals, and how to combine large numbers of possibly conflicting judgements into a single joint annotation. In this paper we focus on the latter challenge. Our aim is to investigate and empirically test methods for aggregating the judgements of large numbers of individuals in a linguistic annotation task conducted via crowdsourcing into a *collective judgement*.

Most researchers who turn to crowdsourcing to collect data use majority voting to combine the participants' responses (Sayeed et al., 2011; Zarcone and Rüd, 2012; Venhuizen et al., 2013). Although in the limit it makes sense to take the judgement of the majority as reflecting the view of the community, in practice we cannot reach out to the full population of speakers, which means that the possible biases amongst the participants we manage to recruit may distort the outcome. Also, given the nature of crowsourcing (rewarding speed rather than quality), some participants may not respond truthfully according to their intuitions as speakers. To address these issues, we propose aggregation methods that go beyond majority voting by taking into account the reliability of individual annotators at the time of aggregation.[1] Our approach is related to existing work on analysing the quality of annotated data by examining, for instance, (dis)agreement patterns amongst annotators (Bhardwaj et al., 2010; Peldszus and Stede, 2013; Ramanath et al., 2013). However, while the main aim of this kind of studies is to gain insight into the difficulty of an annotation task or into the feasibility of using untrainned annotators for particular tasks, our focus is on exploiting patterns of judgements for the purpose of aggregation into a single collective annotation—an aspect that has received far less attention in the literature.

We make the following contributions: (i) we make available two new datasets of judgements gathered with AMT for two multi-category annotation tasks; (ii) we define several aggregation methods based, on

[1] Other aspects can contribute to reduce the shortcomings of crowdsourcing at earlier stages, such as task design and annotator recruiting constraints. However, here we specifically deal with improving quality at the time of aggregation.

the one hand, on an approach by inspired by social choice theory (Endriss and Fernández, 2013; Kruger et al., 2014), and on the other hand, on probabilistic generative models pioneered by Dawid and Skene (1979); and (iii) we systematically evaluate the performance of the proposed methods on three different annotation tasks.[2]

The paper is structured as follows: In the next section, we introduce our aggregation methods. In Section 3, we evaluate their performance on different datasets and analyse the results. We then examine two further aspects: the impact of the number of annotators in Section 4 and the presence of highly unreliable annotators in Section 5. We conclude in Section 6 with plans for future work.

2 Aggregation Methods

In this section, we define several methods for deriving a collective judgement in a linguistic annotation task from a set of individual annotations. We focus on simple classification tasks where collecting these individual annotations via a crowdsourcing platform is feasible.

2.1 Notation and Terminology

In our model, an *annotation task* consists of three finite sets: the *items* J, the *categories* K, and the *annotators* N. Each annotator is asked to label some of the items with a category. A *group annotation* A is an $|N| \times |J|$ matrix, with a_{ij} representing the category $k \in K$ that annotator $i \in N$ assigned to item $j \in J$. Let N_j denote the set of annotators who annotated item j (i.e., a_{ij} is undefined if $i \notin N_j$).

We want to aggregate the information contained in a group annotation into a single *collective annotation* that assigns a category to each item. An *aggregator* is a function F that maps a group annotation A into a collective annotation $F(A)$, a vector of categories with dimensionality $|J|$ labelling every item with a category. The most widely used aggregator is the *simple plurality rule* (SPR)—known as *simple majority* in the two-category case—which returns a collective annotation where each item j is labelled with the category chosen most often for j by the group, i.e., $\text{SPR}(A)_j \in \text{argmax}_{k \in K} |\{i \in N_j \mid a_{ij} = k\}|$. Since the SPR may lead to a tie, if we require a single category for each item, a tie-breaking method (such as random tie-breaking) must be adopted. For the purposes of this paper, we assign the special category *'undecided'* whenever an aggregator produces a tie (this is reasonable also in practice: we would not want to commit to a randomly chosen category for an annotated linguistic resource).

2.2 Frequency-based Aggregation

In previous work, we introduced (Endriss and Fernández, 2013) and further refined (Kruger et al., 2014) a framework for deriving a collective annotation inspired by social choice theory. They propose so-called *bias-correcting rules* (BCR's), which try to take the reliability of annotators into account by considering the *frequencies* with which annotators choose certain categories. For example, if annotator i uses category k very often, then this might be a sign that i is overusing k and we should give her votes for k less weight. However, if k is also a frequent choice of the population of annotators at large, then this might again temper that effect.

For a given group annotation A, define the *individual frequency* of annotator i choosing category k—$\text{Freq}_i(k)$—as the number of times i chooses k, divided by the total number of items she annotates. Define the *global frequency* of k—$\text{Freq}(k)$—as the number of times k is chosen by someone, divided by the total number of individual annotations. Thus, if $\text{Freq}_i(k)$ is high, particularly if $\text{Freq}_i(k) > \text{Freq}(k)$, we may want to give a relatively low weight to any instance of annotator i choosing catgeory k.

Every BCR defines a family of weights w_{ik}, specifying for each annotator $i \in N$ and each category $k \in K$ how much weight to give to i's choice of k:

$$F_w(A)_j \quad \in \quad \operatorname*{argmax}_{k \in K} \sum_{i \in N_j | a_{ij} = k} w_{ik}$$

[2]The new datasets and an implementation of our aggregation methods are available at `http://www.illc.uva.nl/Resources/CollectiveAnnotation/`.

Diff	*difference-based BCR*	$w_{ik} = 1 + \text{Freq}(k) - \text{Freq}_i(k)$	**Rat**	*ratio-based BCR*	$w_{ik} = \text{Freq}(k)/\text{Freq}_i(k)$
Com	*complement-based BCR*	$w_{ik} = 1 + 1/\lvert K\rvert - \text{Freq}_i(k)$	**Inv**	*inverse-based BCR*	$w_{ik} = 1/\text{Freq}_i(k)$

Table 1: Weights used for canonical Bias Correcting Rules.

In case of a tie, we assign category *'undecided'*. Table 1 defines the weights for four specific BCR's. Thus, for example, if an annotator uses k in 50% of the cases, while the general population only uses k in 20% of all cases, then under Diff she has weight 0.7 whenever she chooses k. Note that Com and Inv do not take global frequencies into account, while Diff and Rat do.

2.3 Agreement-based Aggregation

Suppose each item has a *true* (but unknown) category (its *gold standard*). We may view an annotator's judgement as a noisy signal of the gold standard. We now want to design an aggregator as a maximum likelihood estimator for this ground truth. This approach has been pioneered by Dawid and Skene (1979). Variants have been used for diverse purposes by, amongst others, Snow et al. (2008), Carpenter (2008), Raykar et al. (2010), Ipeirotis et al. (2010), Li et al. (2013), and Passonneau and Carpenter (2013).

Let $p(a_{ij} = k \mid g_j = k^\star)$, with k not necessarily distinct from $k^\star$, be the probability of agent $i \in N_j$ annotating item j with category $k \in K$, given that the gold standard category of j is $k^\star \in K$. If we can obtain estimates of these probabilities, then we can use them to calibrate the weights of the annotators. The challenge, particularly for multi-category annotation tasks, is that the number of probabilities to estimate is fairly large (in particular, it is quadratic in $|K|$). To be able to provide reasonable estimates, we need a large amount of data *from every individual annotator*. But this precisely we do not have in crowdsourcing: we have a lot of data, but it comes from many different annotators. We thus make two simplifying assumptions, aimed at aggressively reducing the number of parameters to estimate:[3]

(1) We assume that $p(a_{ij} = k^\star \mid g_j = k^\star)$, i.e., annotator i's probability of choosing the *correct* category, does not depend on either j or $k^\star$. It only depends on i's accuracy. Thus, we can abbreviate $\text{acc}_i := p(a_{ij} = k^\star \mid g_j = k^\star)$.

(2) We assume that when annotator i does not choose the correct category $k^\star$, then she is equally likely to pick any of the *wrong* categories $k \neq k^\star$: $p(a_{ij} = k \mid g_j = k^\star) = \frac{1-\text{acc}_i}{|K|-1}$.

Assumption (1) is not uncommon (Li et al., 2013), but it clearly is a limiting assumption: accuracy not depending on j means that we cannot model the fact that some items are more difficult to label correctly; accuracy not depending on k means that we cannot model the fact that some categories are harder to comprehend than others. Assumption (2) and its alternatives only come into play when there are more than two categories; as large parts of the literature focus on the two-category case, this issue has received less attention. One of the limitations of assumption (2) is that we cannot model that some categories may "look similar" and are likely to get confused with each other.

On the positive side, in our simplified model we only have a single parameter to estimate for each annotator, namely its accuracy acc_i. Now suppose, hypothetically, we knew the acc_i's (which we do not in practice). Which category should we pick for item j? To answer this question we need to consider probabilities such as $p(g_j = k \mid A_j)$, the probability that k is the true category for item j given our observation of column A_j. If we do not want to make any assumptions regarding possible priors for either gold standards or annotation biases (i.e., if we opt for the default assumption of uniform priors), then we can instead work with $p(A_j \mid g_j = k)$. Specifically, we should choose k over k' if $p(A_j \mid g_j = k) > p(A_j \mid g_j = k')$, i.e., if:

$$\prod_{i|a_{ij}=k} \text{acc}_i \prod_{i|a_{ij}=k'} \frac{1-\text{acc}_i}{|K|-1} \prod_{i|a_{ij}\notin\{k,k'\}} \frac{1-\text{acc}_i}{|K|-1} > \prod_{i|a_{ij}=k'} \text{acc}_i \prod_{i|a_{ij}=k} \frac{1-\text{acc}_i}{|K|-1} \prod_{i|a_{ij}\notin\{k,k'\}} \frac{1-\text{acc}_i}{|K|-1}$$

$$\prod_{i|a_{ij}=k} \frac{(|K|-1)\cdot\text{acc}_i}{1-\text{acc}_i} > \prod_{i|a_{ij}=k'} \frac{(|K|-1)\cdot\text{acc}_i}{1-\text{acc}_i}$$

[3] That is, we are trading generality of the model against estimation quality of its parameters (see also Section 3.4).

Taking logarithms on both sides, we see that giving each annotator a weight of $\log \frac{(|K|-1)\cdot \text{acc}_i}{1-\text{acc}_i}$ results in an optimal aggregator. Let us call the corresponding aggregator the *oracle rule* **Ora**. Importantly, this is not a practically useful rule, as in reality we do *not* know the acc_i's. As we shall see, however, it is a useful benchmark, as it allows us to distinguish between loss in quality due to the simplicity of our model and loss in quality accrued during estimation (given that Ora is perfect w.r.t. the latter dimension).[4]

In practice, we need to estimate the acc_i's. We use a particularly simple method and estimate acc_i as i's *agreement* agr_i with the SPR, defined as follows:[5]

$$\text{agr}_i \quad := \quad \frac{|\{j \in J \mid a_{ij} = \text{SPR}(A)_j\}| + 0.5}{|\{j \in J \mid i \text{ annotates } j\}| + 1}$$

We call the rule we obtain using this method, i.e., the rule giving weight $\log \frac{(|K|-1)\cdot \text{agr}_i}{1-\text{agr}_i}$ to annotator i, the *agreement-based rule* **Agr**. There are two natural refinements of Agr one might consider. First, we could attempt to take priors regarding gold standards into account. If $p(k)$ is the prior probability of encountering (true) category k, then we get $p(g_j = k \mid A_j) \propto p(A_j \mid g_j = k) \cdot p(k)$. This corresponds to adding $\log p(k)$ as an extra weight in favour of category k. We can estimate $p(k)$ using either $\text{Freq}(k)$ or the SPR. The second possible refinement is to iterate the process used to estimate acc_i, i.e., to use Agr in place of SPR to compute better estimates agr'_i of acc_i, and so forth. That is, we could use the EM algorithm (Dawid and Skene, 1979) to estimate acc_i. As we shall see, Agr outperforms both of these refinements for the datasets considered in this paper.

3 Performance on Different Datasets

In this section, we evaluate the performance of our aggregation methods on three datasets from three different categorical annotation tasks for which gold standard annotations are readily available. One of these tasks—Recognising Textual Entailment—is a binary classification task and includes non-expert annotations collected by Snow et al. (2008). The other two tasks—Preposition Sense Disambiguation and Question Dialogue Acts—are multi-category tasks for which we have collected new crowdsourced annotations for the purposes of the present study.[6]

3.1 Recognising Textual Entailment (RTE)

This dataset is based on the task proposed by Dagan et al. (2006) in the PASCAL Recognizing Textual Entailment (RTE) Challenge. The RTE task involves deciding whether the meaning of a sentence (the *hypothesis*) can be inferred from a *text*. The original RTE1 Challenge testset consists of 800 text-hypothesis pairs (e.g., T:*"In central Antioquia two ranges of the Colombian Andeas meet"*, H:*"Antioquia is in Colombia."*) with a gold standard annotation that classifies each of them as either *true* (1) or *false* (0), depending on whether H can be inferred from T or not. The released expert annotation is perfectly balanced, with 400 items annotated as 0 and 400 as 1.

Snow et al. (2008) used Amazon's *Mechanical Turk* (AMT) to collect 10 non-expert annotations for each of the 800 items. The annotation task included a total of 164 AMT workers who annotated between 20 items (124 annotators) and 800 items each (only one annotator). Amongst the non-expert annotations, category 1 is slightly more frequent ($\approx 57\%$) than category 0.

Table 2a shows the results of applying the aggregation rules (and the oracle rule) to this data. Here (as later in Tables 2b and 2c), the first columns shows observed agreement (A) between the collective annotation output by each rule and the gold standard.[7] The following columnss show precision and recall for each category. We can see that all rules outperform the SPR.[8] Agr yields better results (93.3%)

[4]Snow et al. (2008) used Dawid and Skene's model to calibrate annotator judgements in terms of the gold standard. In contrast, we only use Ora as a benchmark to get a better understanding of the limitations of our probabilistic model.

[5]The smoothing terms (0.5 and 1) ensure that agr_i will never be 0 or 1, i.e., $\log \frac{(|K|-1)\cdot \text{agr}_i}{1-\text{agr}_i}$ is always well-defined.

[6]For practical reasons, we have opted for evaluating our methods against a gold standard. However, we note that in linguistic tasks, especially those concerning semantics and pragmatics, there may simply not be a 'true' category—a collective annotation may be the closest we can get to representing the view of the community.

[7]All aggregators assign category *'undecided'* in case of a tie. Therefore, any ties are counted as instances of disagreement.

[8]The SPR leads to 65 ties; the other rules lead to none.

	A	0	1
SPR	0.856	.96/.79	.91/.93
Com	0.916	.93/.90	.91/.93
Inv	0.893	.87/.92	.91/.87
Diff	0.915	.94/.88	.89/.95
Rat	0.908	.94/.88	.88/.94
Agr	0.933	.93/.93	.93/.94
[Ora]	0.941	.93/.96	.96/.93

(a) RTE

	A	1	2	3
SPR	0.813	.89/.96	.82/.40	.82/.92
Com	0.820	.87/.95	.70/.46	.82/.92
Inv	0.807	.88/.95	.62/.51	.82/.85
Diff	0.833	.86/.96	.80/.46	.82/.93
Rat	0.840	.87/.96	.81/.49	.82/.93
Agr	0.827	.85/.98	.88/.40	.80/.93
[Ora]	0.833	.85/.98	.88/.43	.81/.93

(b) PSD

	A	1	2	3	4
SPR	0.857	.86/.98	.87/1.0	.92/.75	.90/.42
Com	0.870	.87/.98	.87/1.0	.88/.77	.88/.49
Inv	0.877	.91/.91	.94/.98	.84/.77	.72/.73
Diff	0.867	.84/.98	.87/1.0	.89/.78	.91/.44
Rat	0.870	.84/.99	.87/1.0	.92/.77	.91/.47
Agr	0.867	.84/.99	.87/1.0	.92/.77	.91/.44
[Ora]	0.870	.85/.99	.87/1.0	.92/.77	.91/.47

(c) QDA

Table 2: Observed agreement with the gold standard and precision/recall per category for each task.

than any of the BCR's in this case. For the SPR, category 1 has higher recall than precision, while the opposite is the case for category 0. This is in line with the slightly higher frequency of category 1 in the AMT annotations. The BCR's should be able to correct for this bias and to some extent they do (note the increase in category 0's recall: 88% or higher for any of the BCR's vs. 79% for the SPR). In this dataset, the best-performing BCR is Com (91.6% agreement), keeping a good balance between precision and recall for both categories. If we use the refinement of Agr with priors, then the observed agreement drops slightly (to 92.9% if we estimate gold standard distributions using Freq(k), and to 93.1% if we use the SPR). If we use the EM algorithm to estimate acc_i, the system stabilises after six iterations and the resulting rule also does slightly worse than Agr (93.0%).

3.2 Preposition Sense Disambiguation (PSD)

This annotation task is based on the dataset used in the SemEval 2007 task on word-sense disambiguation of prepositions (Litkowski and Hargraves, 2007). The SemEval dataset consists of roughly 25,000 sentences each containing one of the 34 most common English prepositions. The gold standard annotation was constructed by a single lexicographer who tagged each preposition instance with a sense from the sense inventories given by the Oxford Dictionary of English (ODE).

For our non-expert data collection, we used the 150 sentences with the preposition *among*, which according to ODE has four senses. We simplified the task by collapsing senses 3 and 4, as there is only one item classified with sense 4 by the gold standard and that sense is closest to sense 3.[9] The annotation task was conducted using AMT. We showed the workers the following sense definitions of *among* and asked them to select the appropriate sense for each sentence:

(1) situated more or less centrally in relation to other things, e.g., *"There are flowers hidden among the roots of the trees."*

(2) being a member of a larger set, e.g., *"Snakes are among the animals most feared by man."*

(3) shared by some members of a group or community, e.g., *"Members of the government bickered among themselves."*

The distribution of categories according to the gold standard is 37.3%, 23.3%, and 39.3% for sense 1, 2, and 3, respectively. The non-expert annotation task included 45 AMT workers who annotated between 15 items (26 annotators) and 150 items each (only one annotator; another annotated 135 items). Amongst the AMT annotations, the relative frequency of the categories is 40.6%, 18.8%, and 40.6%, respectively.

The results are shown in Table 2b.[10] The rules with the highest agreement with the gold standard are Diff (83.3%) and Rat (84%), i.e., the rules that take into account the global frequency of the categories. Rat outperforms not only the other three BCR's and the SPR (81.3%) but also Agr (82.7%) and Ora (83.3%). Recall for sense 2 (the rarest category) is low across rules, although less so for the BCR's, which manage to correct slightly for the annotators' bias against this category.[11]

[9] The original ODE sense definitions for *among* can be found at `http://tinyurl.com/ode-among`.

[10] The SPR leads to 6 ties; the other rules lead to none. The two refinements of Agr (priors and EM) do not affect the outcome.

[11] After inspecting the data, we suspect that the gold standard overuses sense 2. For instance, in the folowing sentence *among* is tagged with sense 2 although sense 1 seems more appropriate: "[...] *like icebergs 90 per cent is under the water and that is making them incredibly difficult to see* **among** *the waves."*

3.3 Question Dialogue Acts (QDA)

The second dataset we collected is based on the Switchboard corpus (Godfrey et al., 1992). The corpus includes a gold standard annotation prepared by trained annotators, labelling each utterance with a dialogue act tag from the SWBD-DAMSL annotation scheme (Jurafsky et al., 1997).

For our crowdsourcing experiment, we restricted ourselves to four types of question dialogue acts: Yes-No questions, Wh-questions, Declarative questions (including both declarative wh- and yes-no questions), and Rhetorical questions. We extracted 300 questions from the corpus, 35% of which were annotated as Yes-No in the gold standard, 30% as Wh, 20% as Declarative, and 15% as Rhetorical. The AMT workers were shown the following category definitions (here slightly simplified for space reasons):

(1) Yes-No: Questions with a standard form that could be answered with "yes" or "no" (*"Is that the only pet that you have?"*)

(2) Wh: Questions with a standard form that ask for specific information using wh-words (*"What kind of pet do you have?"*)

(3) Declarative: Questions with a statement-like form that nevertheless ask for an answer (*"You have how many pets."*)

(4) Rhetorical: Questions that do not need to be answered. They can have the form of any of the question types above, but they are asked only to make a point (*"If I ever wanted to have a pet, how could I work?"*)

Each item consists of a short dialogue fragment showing three utterances before and after the question to be annotated. The AMT workers were asked to classify the highlighted question with one of the four question types above. Here is a sample item (with reduced context for space reasons):

A: I understand.
A: **Where is home for you?**
B: Originally, was born in Missouri.

A total of 63 AMT workers participated in the annotation task, annotating between 10 items (24 annotators) and 200 items each (only one annotator). Amongst these non-expert annotations, the relative frequencies for category 1 to 4 are 36.6%, 34.1%, 18.4%, and 10.9%, respectively.

Table 2c shows the results of applying the aggregation rules to this data, plus the outcome of the oracle. [12] Inv yields the best result (87.7%), even outperforming Ora (87%). The annotators tend to overuse the common categories (1 and 2), resulting in high recall but low precision. In contrast, the less frequent categories (3 and 4) tend to be underused, resulting in high precision but low recall. Note how applying Inv leads to particularly high recall for rhetorical questions (category 4). The price to pay is the drop in precision for this category compared to the other rules. The dual effect is that precision for Yes-No (1) and Wh (2) is higher with Inv than with the other rules, while recall is lower.

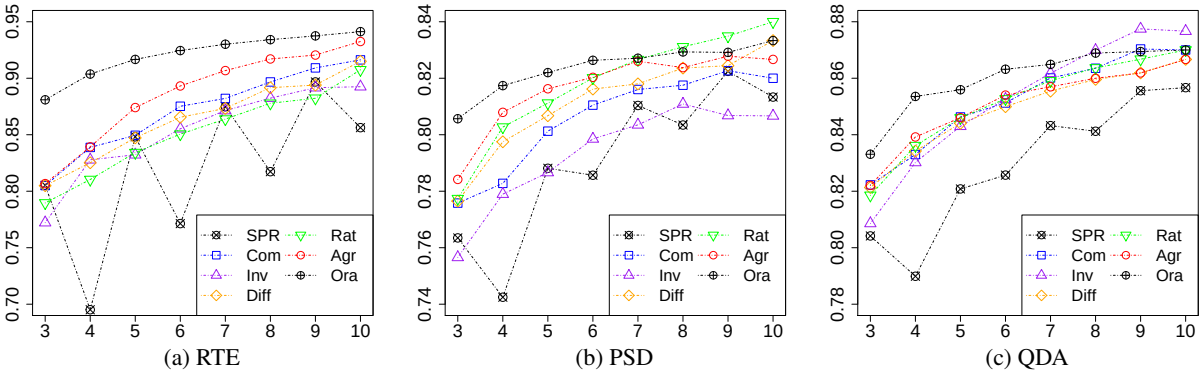
3.4 Comparative Analysis

First, let us compare Agr and Ora. The good performance of Agr suggests that our simple probabilistic model is not *too* simplistic; the trade-off between loss in generality and gain in ability to estimate parameters mentioned in Section 2 appears to be appropriate. The fact that Ora outperforms Agr only slightly suggests that the number of parameters in our model is sufficiently small to be estimated well using the amount of data typically available in linguistic annotation taks conducted via crowdsourcing.

Second, the fact that Agr (modestly) outperforms its refinement using an estimated prior can be explained by the fact that, in our datasets, annotators tend to overuse frequent categories and underuse rare categories. The reason why iterating the rule used to estimate accuracies did not improve performance of Agr for our datasets is less clear, but may be related to the well-known fact that EM can get stuck in a local optimum. The positive take-away message is that the simplest form of our agreement rule resulted in the best performance (at least for our three datasets).

Third, the differences in performance between different BCR's point at an interesting difference in types of bias. Recall that Com and Inv judge the reliability of an annotator only in terms of her own annotations and penalise frequent use of a category. Diff and Rat correct for this effect in case the global frequency is high as well. This means that if a population of annotators has a *shared bias* against or in favour of a category, then Diff and Rat cannot track this well. This explains the fact that Com outperforms Diff and Inv outperforms Rat in the QDA data (see Table 2c): in this task many annotators appeared to

[12] The SPR leads to 7 ties; the other rules to none. Once again, the observed agreement for Agr drops slightly for the two refinements discussed (priors and EM).



(a) RTE (b) PSD (c) QDA

Figure 1: Observed agreement with the gold standard (y-axis) for varying NAI (x-axis).

have difficulties recognising rhetorical questions, i.e., they had a shared bias against labelling an item as Rhetorical. For a dataset with clear *individual biases*, on the other hand, we would expect Diff/Rat to outperform Com/Inv. We do not have a clear case of such a phenomenon in the data analysed here. For the PSD task, Diff/Rat do outperform Com/Inv (see Table 2b), but we believe that the explanation for this finding is a different one: Arguably, the gold standard annotation overuses category 2 (see Footnote 11). This means that high-quality annotators are seen as underusing it and get penalised by Com/Inv. For Diff/Rat this effect is tempered by the fact that the population as a whole is underusing category 2 (relative to the questionable gold standard).

Finally, much can be learned from contraposing the frequency- and agreement-based approach. Suppose the gold standard is uniformly distributed (as for RTE). Then the expected value of $\text{Freq}_i(k)$ is $\frac{1}{|K|}$, i.e., it does not depend on acc_i at all. Thus, the two approaches track entirely different parameters, yet both achieve respectable results. This suggests that combining them might prove fruitful (see Section 5). Certainly, an approach based on a richer probabilistic model would be able to track both kinds of parameters, but as we had argued, this might be infeasible with the relatively small amount of data per annotator we can collect through crowdsourcing. In some sense, what we have done with our rules is trying to make up for the scarcity of data by exploiting our domain knowledge (e.g., regarding the relationships between observed frequency and annotator reliability) to reduce the parameter space.

4 Impact of Number of Annotators

The cost and quality of an annotated linguistic resource created via crowdsourcing crucially depends on the number of annotators that label each item. Having low numbers of coders will make the task more affordable (in terms of time and money), but it will also make the aggregation process more vulnerable to low-quality annotators. Snow et al. (2008) showed how the number of annotators per item (henceforth NAI) influences the performance of the SPR. Here we further explore the impact of NAI on the quality of the collective annotation obtained by different aggregation methods.

For each of the three datasets and each NAI n ($3 \leqslant n \leqslant 9$), we randomly resampled n annotations for each set of items presented to a worker in one go (i.e., for each HIT in AMT terminology). This allowed us to generate a subset of the original dataset with n annotators per item. We generated 1000 such random subsets for each n, applied our aggregators to each subset (and also computed the oracle outcome). We then calculated the average observed agreement with the gold standard. To test whether the differences observed are statistically significant, we calculated the difference in performance between pairs of rules on each subset and computed the 95% (one-sided) confidence intervals by using its distribution over the 1000 subsets. If the proportion of subsets on which this difference is strictly greater than 0 is higher than 95%, we consider the difference to be significant.

The results are shown in Figure 1. We can see that, as the NAI increases, the performance of the rules generally improves (except for the oscillation of the SPR due to tie-breaking). This improvement is greater when the NAI is small (from 3 to 5), which suggest that a minimum of 5 annotators per item is

	0	6
SPR	0.856	0.911
Com	0.916	0.930
Inv	0.893	0.933
Diff	0.915	0.928
Rat	0.908	0.926
Agr	0.933	0.929
[Ora]	0.941	0.944

(a) RTE

	0	9
SPR	0.813	0.820
Com	0.820	0.840
Inv	0.807	0.840
Diff	0.833	0.820
Rat	0.840	0.833
Agr	0.827	0.827
[Ora]	0.833	0.827

(b) PSD

	0	6
SPR	0.857	0.867
Com	0.870	0.883
Inv	0.877	0.903
Diff	0.867	0.873
Rat	0.870	0.877
Agr	0.867	0.867
[Ora]	0.870	0.883

(c) QDA

Table 3: Effect on observed agreement when removing 6 spammers in RTE, 9 in PSD, and 6 in QDA.

recommended. We can also observe that Agr has a robust performance on all datasets when the NAI is between 5 and 7: its improvement over the SPR is statistically significant in all cases for the three tasks, except on PSD when NAI is 7, in which case it is neither significantly better nor significantly worse than the SPR. Note that in all datasets Agr only needs 6 or 7 annotators per item to achieve an accuracy comparable to the SPR using 10 annotators per item.

The robustness of Agr with low NAI is not surprising, given that it already assigns low weights to workers who consistently disagree with the majority. Discounting such problematic workers is particularly important when there are relatively few workers per item. But as the NAI increases, it becomes more likely that random annotators will cancel each other out. It is then that we observe the greatest advantage of using BCR's. This can be seen in the plots for PSD and QDA with high NAI. In those cases the improvement of the best performing BCR's (Rat on PSD and Inv on QDA) over the other rules approaches significance although does not reach the 95% threshold (e.g., on QDA when the NAI is 9, Inv is strictly better than Agr for 93.4% of the subsets).

5 Removal of Low-Quality Annotators

Next we discuss how removing easily recognisable low-quality annotators ("spammers") before aggregation affects the quality of results. The BCR's make the implicit assumption that annotators are sincere. This can be problematic, given the nature of crowdsourcing, where it is not uncommon to encounter workers giving random rather than truthful responses (Sheng et al., 2008; Raykar and Yu, 2012). BCR's are vulnerable to this phenomenon. Here we propose to combine the frequency- and agreement-based approach by using the agreement rate of an annotator with the SPR outcome to identfy and remove spammers prior to applying the frequency-based BCR's.

We take *spammers* to be those annotators that annotate a large number of items (i.e., we have sufficient evidence to judge) and that systematically deviate from the plurality outcome. In the specific context of our datasets, we have implemented this idea by labelling as spammers those annotators who annotated at least 20% of the total number of items and whose agreement rate with the SPR is below the median agreement rate. This corresponds to 6 annotators in the RTE dataset, 9 in the PSD dataset, and 6 in the QDA dataset. The effect of removing these low-quality annotators from the population can be seen in Table 3 showing observed agreement of the different aggregation rules (and the oracle rule) with the gold standard before and after spammer removal.

The results show that, with one exception, after removing spammers the performance of the BCR's improves significantly. The exception concerns Diff and Rat for the PSD dataset. Recall that the gold standard for this dataset, arguably, overuses category 2 (see Footnote 11 and Section 3.4). That is, high-quality annotators are (wrongly) judged to be underusing category 2. Before spammer removal, this effect is tempered by the presence of a few annotators delivering 'random' annotations (thereby artificially increasing the frequency of category 2). After spammer removal, this positive effect is diminished and rules such as Diff and Rat suffer in performance. Con and Inv, on the other hand, can compensate for this effect simply by giving very high weights to those (high-quality) annotators who still use the relatively rare category 2. Also for RTE and QDA, amongst the BCR's the rules not based on global frequencies, i.e., Com and Inv, benefit most. Indeed, after spammer removal Com/Inv perform better than Diff/Rat for all three datasets. Overall, Inv with spammer removal is our best-performing rule.

Not surprisingly, Agr and Ora gain relatively little from spammer removal since, given our definition of a spammer, the removed annotators already had very low weights to begin with. In fact, the performance of these aggregation rules may even drop slightly after removing spammers (see Tables 3a and 3b).

6 Conclusions

We have argued that simply using the majority/plurality rule to aggregate individual linguistic judgments in a crowdsourcing annotation task is far from optimal. Instead, we have proposed several methods that weight the annotators' judgements by exploiting either the frequency with which they choose particular categories or the degree to which they agree with the full population of annotators. We have tested our methods on existing datasets and we have also created two new datasets. Our results show how annotation tasks with different characteristics can benefit from different types of aggregation methods. Our aggregation methods result in small but robust gains across datasets, both in terms of accuracy achieved and in terms of the number of annotators required to obtain acceptable results.

Besides BCR's, in our previous work we also proposed a *greedy consensus rule*, albeit only for the two-category case (Endriss and Fernández, 2013) . This rule sequentially locks in simple majorities in the order of relative majority strength, but along the way disregards annotators who disagree with too many of those strong majorities. It performs well on the RTE dataset (almost as well as Agr). Intuitively speaking, it can track *item difficulty*, by first settling the easy items (with clear majorities) and thereby learning which annotators are most reliable to then have them decide on the harder items. Here we have not included this rule as there is no single most natural way of generalising it to the multi-category case. Arriving at such a generalisation in a principled manner is an important direction for future work.

It would also be interesting to get a clearer understanding of the links between methods for assessing inter-annotator agreement (Artstein and Poesio, 2008) and methods of aggregation (i.e., methods that may be applied to data of possibly rather poor inter-annotator agreement, as is the case for parts of our datasets). A relevant observation in this context is that the notions of individual and global frequency at the core of our BCR's also play a role in agreement coefficients, namely to compute chance agreement: π (Scott, 1955) uses global frequencies and κ (Cohen, 1960) uses individual frequencies.

While the definition of Agr was motivated by a simple probabilistic model, the BCR's were motivated by rules of thumb regarding links between observed frequencies and reliability. We have noted before that the BCR's do not track the same phenomena as Agr; rather, they seem to complement each other, an observation we have exploited explicitly when removing spammers before applying a BCR. Identifying a suitable probabilistic model for our frequency-based BCR's promises to be a fruitful future line of research, as it would allow for a better comparison (and eventually integration) of the two approaches.

References

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

Vikas Bhardwaj, Rebecca J Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: a framework for analysis of multiple annotators' labeling behavior. In *Proc. 4th Linguistic Annotation Workshop*, pages 47–55. ACL.

Bob Carpenter. 2008. Multilevel Bayesian Models of Categorical Data Annotation. Technical report, LingPipe.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, volume 3944 of *LNCS*, pages 177–190. Springer-Verlag.

Alexander P. Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28.

Ulle Endriss and Raquel Fernández. 2013. Collective annotation of linguistic resources: Basic principles and a formal model. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, pages 539–549.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Devlopment. In *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 517–520.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proc. 2nd Human Computation Workshop (HCOMP-2010)*.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function-annotation coder's manual, draft 13. Technical Report TR 97-02, Institute for Cognitive Science, University of Colorado at Boulder.

Justin Kruger, Ulle Endriss, Raquel Fernández, and Ciyang Qing. 2014. Axiomatic analysis of aggregation methods for collective annotation. In *Proc. 13th Int'l Conference on Autonomous Agents and Multiagent Systems (AAMAS-2014)*, pages 1185–1192. IFAAMAS.

Hongwei Li, Bin Yu, and Dengyong Zhou. 2013. Error rate analysis of labeling by crowdsourcing. In *Proc. Machine Learning meets Crowdsourcing, Workshop at the Int'l Conference on Machine Learning (ICML-2013)*.

Kenneth C. Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proc. 4th Int'l Workshop on Semantic Evaluations (SemEval-2007)*.

Rebecca J. Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. In *Proc. 7th Linguistic Annotation Workshop*, pages 187–195. ACL.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proc. 7th Linguistic Annotation Workshop*, pages 196–204. ACL.

Rohan Ramanath, Monojit Choudhury, Kalika Bali, and Rishiraj Saha Roy. 2013. Crowd prefers the middle path: A new iaa metric for crowdsourcing reveals turker biases in query segmentation. *Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, pages 1713–1722.

Vikas Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13:491–518.

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

Asad Sayeed, Bryan Rusk, Martin Petrov, Hieu Nguyen, Timothy Meyer, and Amy Weinber. 2011. Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proc. Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH-2011)*.

William A. Scott. 1955. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quaterly*, 19(3):321–325.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. 14th ACM Int'l Conference on Knowledge Discovery and Data Mining (KDD-208)*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 254–263.

Noortje Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for word sense labeling. In *Proc. 10th Int'l Conference on Computational Semantics (IWCS-2013)*, pages 397–403.

Alessandra Zarcone and Stefan Rüd. 2012. Logical metonymies and qualia structures: An annotated database of logical metonymies for German. In *Proc. Language Resources and Evaluation Conference (LREC-2012)*, pages 1799–1804.