Collective Annotation of Linguistic Resources: Basic Principles and a Formal Model

Ulle Endriss and **Raquel Fernández**
Institute for Logic, Language & Computation
University of Amsterdam
{ulle.endriss|raquel.fernandez}@uva.nl

Abstract

Crowdsourcing, which offers new ways of cheaply and quickly gathering large amounts of information contributed by volunteers online, has revolutionised the collection of labelled data. Yet, to create annotated linguistic resources from this data, we face the challenge of having to combine the judgements of a potentially large group of annotators. In this paper we investigate how to aggregate individual annotations into a single collective annotation, taking inspiration from the field of social choice theory. We formulate a general formal model for collective annotation and propose several aggregation methods that go beyond the commonly used majority rule. We test some of our methods on data from a crowdsourcing experiment on textual entailment annotation.

1 Introduction

In recent years, the possibility to undertake large-scale annotation projects with hundreds or thousands of annotators has become a reality thanks to online crowdsourcing methods such as Amazon's *Mechanical Turk* and *Games with a Purpose*. Although these techniques open the door to a true revolution for the creation of annotated corpora, within the computational linguistics community there so far is no clear understanding of how the so-called "wisdom of the crowds" could or should be used to develop useful annotated linguistic resources. Those who have looked into this increasingly important issue have mostly concentrated on validating the quality of multiple non-expert annotations in terms of how they compare to expert gold standards; but they have only used simple aggregation methods based on majority voting to combine the judgments of individual annotators (Snow et al., 2008; Venhuizen et al., 2013).

In this paper, we take a different perspective and instead focus on investigating different aggregation methods for deriving a single *collective annotation* from a diverse set of judgments. For this we draw inspiration from the field of *social choice theory*, a theoretical framework for combining the preferences or choices of several individuals into a collective decision (Arrow et al., 2002). Our aim is to explore the parallels between the task of aggregating the preferences of the citizens participating in an election and the task of combining the expertise of speakers taking part in an annotation project. Our contribution consists in the formulation of a general formal model for collective annotation and, in particular, the introduction of several families of aggregation methods that go beyond the commonly used majority rule.

The remainder of this paper is organised as follows. In Section 2 we introduce some basic terminology and argue that there are four natural forms of collective annotation. We then focus on one of them and present a formal model for it in Section 3. We also formulate some basic principles of aggregation within this model in the same section. Section 4 introduces three families of aggregation methods: bias-correcting majority rules, greedy methods for identifying (near-)consensual coalitions of annotators, and distance-based aggregators. We test the former two families of aggregators, as well as the simple majority rule commonly used in similar studies, in a case study on data extracted from a crowdsourcing experiment on textual entailment in Section 5. Section 6 discusses related work and Section 7 concludes.

2 Four Types of Collective Annotation

An annotation task consists of a set of *items*, each of which is associated with a set of possible *categories* (Artstein and Poesio, 2008). The categories may be the same for all items or they may be item-specific. For instance, dialogue act annotation

(Allen and Core, 1997; Carletta et al., 1997) and word similarity rating (Miller and Charles, 1991; Finkelstein et al., 2002) involve choosing from amongst a set of categories—acts in a dialogue act taxonomy or values on a scale, respectively—which remains fixed for all items in the annotation task. In contrast, in tasks such as word sense labelling (Kilgarriff and Palmer, 2000; Palmer et al., 2007; Venhuizen et al., 2013) and PP-attachment annotation (Rosenthal et al., 2010; Jha et al., 2010) coders need to choose a category amongst a set of options specific to each item—the possible senses of each word or the possible attachment points in each sentence with a prepositional phrase.

In either case (one set of categories for all items *vs.* item-specific sets of categories), annotators are typically asked to identify, for each item, the category they consider the best match. In addition, they may be given the opportunity to indicate that they cannot judge (the "don't know" or "unclear" category). For large-scale annotation projects run over the Internet it is furthermore very likely that an annotator will not be confronted with every single item, and it makes sense to distinguish items not seen by the annotator from items labelled as "don't know". We refer to this form of annotation, i.e., an annotation task where coders have the option to (i) label items with one of the available categories, to (ii) choose "don't know", or to (iii) not label an item at all, as *plain annotation*.

Plain annotation is the most common form of annotation and it is the one we shall focus on in this paper. However, other, more complex, forms of annotation are also possible and of interest. For instance, we may ask coders to *rank* the available categories (resulting in, say, a weak or partial order over the categories); we may ask them to provide a *qualitative ratings* of the available categories for each item (e.g., *excellent match*, *good match*, etc.); or we may ask for *quantitative ratings* (e.g., numbers from 1 to 100).[1] We refer to these forms of annotation as *complex annotation*.

We want to investigate how to aggregate the information available for each item once annotations by multiple annotators have been collected. In line with the terminology used in social choice theory and particularly judgment aggregation (Arrow, 1963; List and Pettit, 2002), let us call an aggregation method *independent* if the outcome regarding a given item j only depends on the categories provided by the annotators regarding j itself (but not on, say, the categories assigned to a different item j'). Independent aggregation methods are attractive due to their simplicity. They also have some conceptual appeal: when deciding on j maybe we *should* only concern ourselves with what people have to say regarding j? On the other hand, insisting on independence prevents us from exploiting potentially useful information that cuts across items. For instance, if a particular annotator almost always chooses category c, then we should maybe give less weight to her selecting c for the item j at hand than when some other annotator chooses c for j. This would call for methods that do not respect independence, which we shall refer to as *general* aggregation. Note that when studying independent aggregation methods, without loss of generality, we may assume that each annotation task consists of just a single item.

In view of our discussion above, there are four classes of approaches to collective annotation:

(1) *Independent aggregation of plain annotations.* This is the simplest case, resulting in a fairly limited design space. When, for a given item, each annotator has to choose between k categories (or abstain) and we do not permit ourselves to use any other information, then the only reasonable choice is to implement the *plurality rule* (Taylor, 2005), under which the winning category is the category chosen by the largest number of annotators. In case there are exactly two categories available, the plurality rule is also called the *majority rule*. The only additional consideration to make here (besides how to deal with ties) is whether or not we may want to declare no winner at all in case the plurality winner does not win by a sufficiently significant margin or does not make a particular quota. This is the most common approach in the literature (see, e.g., Venhuizen et al., 2013).

(2) *Independent aggregation of complex annotations.* This is a natural generalisation of the first approach, resulting in a wider range of possible methods. We shall not explore it here, but only point out that in case annotators provide *linear orders* over categories, there is a close resemblance to classical voting the-

[1] Some authors have combined qualitative and quantitative ratings; e.g., for the Graded Word Sense dataset of Erk et al. (2009) coders were asked to classify each relevant WordNet sense for a given item on a 5-point scale: 1 *completely different*, 2 *mostly different*, 3 *similar*, 4 *very similar*, 5 *identical*.

ory (Taylor, 2005); in case only *partial orders* can be elicited, recent work in computational social choice on the generalisation of classical voting rules may prove helpful (Pini et al., 2009; Endriss et al., 2009); and in case annotators rate categories using qualitative expressions such as *excellent match*, the method of *majority judgment* of Balinski and Laraki (2011) should be considered.

(3) *General aggregation of plain annotations.* This is the approach we shall discuss below. It is related to *voting in combinatorial domains* studied in computational social choice (Chevaleyre et al., 2008), and to both *binary aggregation* (Dokow and Holzman, 2010; Grandi and Endriss, 2011) and *judgment aggregation* (List and Pettit, 2002).

(4) *General aggregation of complex annotations.* While appealing due to its great level of generality, this approach can only be tackled successfully once approaches (2) and (3) are sufficiently well understood.

3 Formal Model

Next we present our model for general aggregation of plain annotations into a collective annotation.

3.1 Terminology and Notation

An *annotation task* is defined in terms of m *items*, with each item $j \in \{1, \dots, m\}$ being associated with a finite set of possible *categories* $\mathcal{C}_j$. Annotators are asked to provide an answer for each of the items of the annotation task. In the context of plain annotations, a valid *answer* for item j is an element of the set $\mathcal{A}_j = \mathcal{C}_j \cup \{?, \bot\}$.[2] Here ? represents the answer "don't know" and we use $\bot$ to indicate that the annotator has not answered (or even seen) the item at all. An *annotation* is a vector of answers by one annotator, one answer for each item of the annotation task at hand, i.e., an annotation is an element of the Cartesian product $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_m$. A typical element of $\mathcal{A}$ will be denoted as $A = (a_1, \dots, a_m)$.

Let $\mathcal{N} = \{1, \dots, n\}$ be a finite set of n *annotators* (or *coders*). A *profile* $\boldsymbol{A} = (A_1, \dots, A_n) \in \mathcal{A}^n$, for a given annotation task, is a vector of annotations, one for each annotator. That is, $\boldsymbol{A}$ is an $n \times m$-matrix; e.g., $a_{3,7}$ is the answer that the 3rd annotator provides for the 7th item.

	Item 1	Item 2	Item 3
Annotator 1	B	A	A
Annotator 2	B	B	B
Annotator 3	A	B	A
Majority	B	B	A

Table 1: A profile with a collective annotation.

We want to aggregate the information provided by the annotators into a (single) collective annotation. For the sake of simplicity, we use $\mathcal{A}$ also as the domain of possible collective annotations (even though the distinction between ? and $\bot$ may not be strictly needed here; they both indicate that we do not want to commit to any particular category). An *aggregator* is a function $F : \mathcal{A}^n \to \mathcal{A}$, mapping any given profile into a collective annotation, i.e., a labelling of the items in the annotation task with corresponding categories (or ? or $\bot$). An example is the *plurality rule* (also known as the *majority rule* for binary tasks with $|\mathcal{C}_j| = 2$ for all items j), which annotates each item with the category chosen most often.

Note that the collective annotation need not coincide with any of the individual annotations. Take, for example, a binary annotation task in which three coders label three items with category A or B as shown in Table 1. Here using the majority rule to aggregate the annotations would result in a collective annotation that does not fully match any annotation by an individual coder.

3.2 Basic Properties

A typical task in social choice theory is to formulate *axioms* that formalise specific desirable properties of an aggregator F (Arrow et al., 2002). Below we adapt three of the most basic axioms that have been considered in the social choice literature to our setting and we briefly discuss their relevance to collective annotation tasks.

We will require some additional notation: for any profile $\boldsymbol{A}$, item j, and possible answer $a \in \mathcal{A}_j$, let $N^{\boldsymbol{A}}_{j:a}$ denote the set of annotators who chose answer a for item j under profile $\boldsymbol{A}$.

- F is *anonymous* if it treats coders symmetrically, i.e., if for every permutation $\pi : \mathcal{N} \to \mathcal{N}$, $F(A_1, \dots, A_n) = F(A_{\pi(1)}, \dots, A_{\pi(n)})$. In social choice theory, this is a fairness constraint. For us, fairness *per se* is not a desideratum,

[2] As discussed earlier, in the context of *complex* annotations, an answer could also be, say, a partial order on $\mathcal{C}_j$ or a function associating elements of $\mathcal{C}_j$ with numerical ratings.

but when we do not have any *a priori* information regarding the expertise of annotators, then anonymity is a natural axiom to adopt.

- F is *neutral* if it treats all items symmetrically, i.e., if for every two items j and j' with the same set of possible categories (i.e., with $\mathcal{C}_j = \mathcal{C}_{j'}$) and for every profile $\boldsymbol{A}$, it is the case that whenever $N^{\boldsymbol{A}}_{j:a} = N^{\boldsymbol{A}}_{j':a}$ for all answers $a \in \mathcal{A}_j = \mathcal{A}_{j'}$, then $F(\boldsymbol{A})_j = F(\boldsymbol{A})_{j'}$. That is, if the patterns of individual annotations of j and j' are the same, then also their collective annotation should coincide. In social choice theory, neutrality is also considered a basic fairness requirement (avoiding preferential treatment one candidate in an election). In the context of collective annotation there may be good reasons to violate neutrality: e.g., we may use an aggregator that assigns different default categories to different items and that can override such a default decision only in the presence of a significant majority (note that this is different from anonymity: we will often not have any information on our annotators, but we may have tangible information on items).[3]
- F is *independent* if the collective annotation of any given item j only depends on the individual annotations of j. Formally, F is independent if, for every item j and every two profiles $\boldsymbol{A}$ and $\boldsymbol{A'}$, it is the case that whenever $N^{\boldsymbol{A}}_{j:a} = N^{\boldsymbol{A'}}_{j:a}$ for all answers $a \in \mathcal{A}_j$, then $F(\boldsymbol{A})_j = F(\boldsymbol{A'})_j$. In social choice theory, independence is often seen as a desirable albeit hard (or even impossible) to achieve property (Arrow, 1963). For collective annotation, we strongly believe that it is *not* a desirable property: by considering how annotators label other items we can learn about their biases and we should try to exploit this information to obtain the best possible annotation for the item at hand.

Note that the plurality/majority rule is independent. All of the methods we shall propose in Section 4 are both anonymous and neutral—except to the extent to which we have to violate basic symmetry requirements in order to break ties between categories chosen equally often for a given item. None of our aggregators is independent.

Some annotation tasks might be subject to *integrity constraints* that determine the internal consistency of an annotation. For example, if our items are pairs of words and the possible categories include *synonymous* and *antonymous*, then if item 1 is about words A and B, item 2 about words B and C, and item 3 about words A and C, then any annotation that labels items 1 and 2 as *synonymous* should not label item 3 as *antonymous*. Thus, a further desirable property that will play a role for some annotation tasks is *collective rationality* (Grandi and Endriss, 2011): if all individual annotations respect a given integrity constraint, then so should the collective annotation.

We can think of integrity constraints as imposing top-down expert knowledge on an annotation. However, for some annotation tasks, no integrity constraints may be known to us in advance, even though we may have reasons to believe that the individual annotators do respect some such constraints. In that case, selecting one of the individual annotations in the profile as the collective annotation is the only way to ensure that these integrity constraints will be satisfied by the collective annotation (Grandi and Endriss, 2011). Of course, to do so we would need to assume that there is at least one annotator who has labelled *all* items (and to be able to design a high-quality aggregator in this way we should have a sufficiently large number of such annotators to choose from), which may not always be possible, particularly in the context of crowdsourcing.

4 Three Families of Aggregators

In this section we instantiate our formal model by proposing three families of methods for aggregation. Each of them is inspired, in part, by standard approaches to desigining aggregation rules developed in social choice theory and, in part, by the specific needs of collective annotation. Regarding the latter point, we specifically emphasise the fact that not all annotators can be expected to be equally reliable (in general or w.r.t. certain items) and we try to integrate the process of aggregation with a process whereby less reliable annotators are either given less weight or are excluded altogether.

4.1 Bias-Correcting Majority Rules

We first want to explore the following idea: If a given annotator annotates *most* items with 0, then we might want to assign less significance to that

[3] It would also be of interest to formulate a neutrality axiom w.r.t. categories (rather than items). For two categories, this idea has been discussed under the name of *domain-neutrality* in the literature (Grandi and Endriss, 2011), but for larger sets of categories it has not yet been explored.

choice for any particular item.[4] That is, if an annotator appears to be biased towards a particular category, then we might want to try to correct for this bias during aggregation.

What follows applies only to annotation tasks where every item is associated with the same set of categories. For ease of exposition, let us furthermore assume that there are only two categories, 0 and 1, and that annotators do not make use of the option to annotate with ? ("don't know").

For every annotator $i \in \mathcal{N}$ and every category $X \in \{0, 1\}$, fix a *weight* $w_i^X \in \mathbb{R}$. The *bias-correcting majority* (BCM) rule for this family of weights is defined as follows. Given profile $\boldsymbol{A}$, the collective category for item j will be 1 in case $\sum_{a_{i,j}=1} w_i^1 > \sum_{a_{i,j}=0} w_i^0$, and 0 otherwise.[5] That is, we compute the overall weight for category 1 by adding up the corresponding weights for those coders that chose 1 for item j, and we do accordingly for the overall weight for category 0; finally, we choose as collective category that category with the larger overall weight. Note that for $w_i^X \equiv 1$ we obtain the simple majority rule.

Below we define three intuitively appealing families of weights, and thereby three BCM rules. However, before we do so, we first require some additional notation. Fix a profile of annotations. For $X \in \{0, 1\}$, let $\mathit{Freq}_i(X)$ denote the relative frequency with which annotator i has chosen category X. For instance, if i has annotated 20 items and has chosen 1 in five cases, then $\mathit{Freq}_i(1) = 0.25$. Similarly, let $\mathit{Freq}(X)$ denote the frequency of X across the entire profile.

Here are three ways of making the intuitive idea of bias correction concrete:

(1) The *complement-based* BCM rule (ComBCM) is defined by weights $w_i^X = \mathit{Freq}_i(1{-}X)$. That is, the weight of annotator i for category X is equal to her relative frequency of having chosen the other category $1{-}X$. For example, if you annotate two items with 1 and eight with 0, then each of your 1-annotations will have weight 0.8, while each of your 0-annotations will only have weight 0.2.

(2) The *difference-based* BCM rule (DiffBCM) is defined by weights $w_i^X = 1 + \mathit{Freq}(X) - \mathit{Freq}_i(X)$. Recall that $\mathit{Freq}(X)$ is the relative frequency of X in the entire profile, while $\mathit{Freq}_i(X)$ is the relative frequency of X in the annotation of i. Hence, if i assigns category X less often than the general population, then her weight on X-choices will be increased by the difference (and *vice versa* in case she assigns X more often than the population at large). For example, if you assign 1 in two out of ten cases, while in general category 1 appears in exactly 50% of all annotations, then your weight for a choice of 1 will be $1 + 0.5 - 0.2 = 1.3$, while you weight for a choice of 0 will only be 0.7.

(3) The *relative* BCM rule (RelBCM) is defined by weights $w_i^X = \frac{\mathit{Freq}(X)}{\mathit{Freq}_i(X)}$. The idea is very similar to the DiffBCM rule. For the example given above, your weight for a choice of 1 would be $0.5/0.2 = 2.5$, while your weight for a choice of 0 would be $0.5/0.8 = 0.625$.

The main difference between the ComBCM rule and the other two rules is that the former only takes into account the possible bias of individual annotators, while the latter two factor in as well the possible skewness of the data (as reflected by the labelling behaviour of the full set of annotators).

In addition, while ComBCM is specific to the case of two categories, DiffBCM and RelBCM immediately generalise to any number of categories. In this case, we add up the category-specific weights as before and then choose the category with maximal support (i.e., we generalise the majority rule underlying the family of BCM rules to the plurality rule).

We stress that our bias-correcting majority rules do *not* violate anonymity (nor neutrality for that matter). If we were to give less weight to a given annotator based on, say, her name, this would constitute a violation of anonymity; if we do so due to properties of the profile at hand and if we do so in a symmetric manner, then it does not.

4.2 Greedy Consensus Rules

Now consider the following idea: If for a given item there is almost complete consensus amongst those coders that annotated it with a proper category (i.e., those who did not choose ? or $\perp$), then we should probably adopt their choice for the collective annotation. Indeed, most aggregators will make this recommendation. Furthermore, the fact that there is almost full consensus for one item

[4] A similar idea is at the heart of *cumulative voting*, which requires a voter to distribute a fixed number of points amongst the candidates (Glasser, 1959; Brams and Fishburn, 2002).

[5] For the sake of simplicity, our description here presupposes that ties are always broken in favour of 0. Other tie-breaking rules (e.g., random tie-breaking) are possible.

may cast doubts on the reliability of coders who disagree with this near-consensus choice and we might want to disregard their views not only w.r.t. that item but also as far as the annotation of other items is concerned. Next we propose a family of aggregators that implement this idea.

For simplicity, suppose that the only proper categories available are 0 and 1 and that annotators do not make use of ? (but it is easy to generalise to arbitrary numbers of categories and scenarios where different items are associated with different categories). Fix a *tolerance value* $t \in \{0, \dots, m\}$. The *greedy consensus rule* GreedyCR^t works as follows. First, initialise the set $\mathcal{N}^\star$ with the full population of annotators $\mathcal{N}$. Then iterate the following two steps:

(1) Find the item with the strongest majority for either 0 or 1 amongst coders in $\mathcal{N}^\star$ and lock in that value for the collective annotation.

(2) Eliminate all coders from $\mathcal{N}^\star$ who disagree on more than t items with the values locked in for the collective annotation so far.

Repeat this process until the categories for all m items have been settled.[6] We may think of this as a "greedy" way of identifying a coalition $\mathcal{N}^\star$ with high inter-annotator agreement and then applying the majority rule to this coalition to obtain the collective annotation.

To be precise, the above is a description of an entire *family* of aggregators: Whenever there is more than one item with a majority of maximal strength, we could choose to lock in any one of them. Also, when there is a split majority between annotators in $\mathcal{N}^\star$ voting 0 and those voting 1, we have to use a tie-breaking rule to make a decision. Additional heuristics may be used to make these local decisions, or they may be left to chance.

Note that in case $t = m$, GreedyCR^t is simply the majority rule (as no annotator will ever get eliminated). In case $t = 0$, we end up with a coalition of annotators that unanimously agree with all of the categories chosen for the collective annotation. However, this coalition of perfectly aligned annotators need not be the largest such coalition (due to the greedy nature of our rule).

Note that greedy consensus rules, as defined here, are both anonymous and neutral. Specifically, it is important not to confuse possible skewness of the data with a violation of neutrality of the aggregator.

4.3 Distance-based Aggregation

Our third approach is based on the notion of *distance*. We first define a metric on choices to be able to say how distant two choices are. This induces an aggregator that, for a given profile, returns a collective choice that minimises the sum of distances to the individual choices in the profile.[7] This opens up a wide range of possibilities; we only sketch some of them here.

A natural choice is the *adjusted Hamming distance* $H : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_{\geqslant 0}$, which counts how many items two annotations differ on:

$$H(A, A') \;=\; \sum_{j=1}^{m} \delta(a_j, a'_j)$$

Here δ is the *adjusted discrete distance* defined as $\delta(x, y) = 0$ if $x = y$ or $x \in \{?, \bot\}$ or $y \in \{?, \bot\}$, and as $\delta(x, y) = 1$ in all other cases.[8]

Once we have fixed a distance d on $\mathcal{A}$ (such as H), this induces an aggregator F_d:

$$F_d(\boldsymbol{A}) \;=\; \underset{A \in \mathcal{A}}{\operatorname{argmin}} \sum_{i=1}^{n} d(A, A_i)$$

To be precise, F_d is an *irresolute* aggregator that might return a *set* of best annotations with minimal distance to the profile.

Note that F_H is simply the plurality rule. This is so because every element of the Cartesian product is a possible annotation. In the presence of integrity constraints excluding some combinations, however, a distance-based rule allows for more sophisticated forms of aggregation (by choosing the optimal annotation w.r.t. all feasible annotations).

We may also try to restrict the computation of distances to a subset of "reliable" annotators. Consider the following idea: If a group of annotators is (fairly) reliable, then they should have a

[6] There are some similarities to Tideman's *Ranked Pairs* method for preference aggregation (Tideman, 1987), which works by fixing the relative rankings of pairs of alternatives in order of the strength of the supporting majorities. In preference aggregation (unlike here), the population of voters is *not* reduced in the process; instead, decisions against the majority are taken whenever this is necessary to guarantee the transitivity of the resulting collective preference order.

[7] This idea has been used in voting (Kemeny, 1959), belief merging (Konieczny and Pino Pérez, 2002), and judgment aggregation (Miller and Osherson, 2009).

[8] This δ, divided by m, is the same thing as what Artstein and Poesio (2008) call the agreement value agr_j for item j.

(fairly) high inter-annotator agreement. By this reasoning, we should choose a group of annotators $\textsc{Ann} \subseteq \mathcal{N}$ that maximises inter-annotator agreement in $\textsc{Ann}$ and work with the aggregator $\mathrm{argmin}_{A \in \mathcal{A}} \sum_{i \in \textsc{Ann}} d(A, A_i)$. But this is too simplistic: any singleton $\textsc{Ann} = \{i\}$ will result in perfect agreement. That is, while we can easily maximise agreement, doing so in a naïve way means ignoring most of the information collected. In other words, we face the following dilemma:

- On the one hand, we should choose a *small* set $\textsc{Ann}$ (i.e., select *few* annotators to base our collective annotation on), as that will allow us to increase the (average) reliability of the annotators taken into account.
- On the other hand, we should choose a *large* set $\textsc{Ann}$ (i.e., select *many* annotators to base our collective annotation on), as that will increase the amount of information exploited.

One pragmatic approach is to fix a minimum quality threshold regarding one of the two dimensions and optimise in view of the other.[9]

5 A Case Study

In this section, we report on a case study in which we have tested our bias-correcting majority and greedy consensus rules.[10] We have used the dataset created by Snow et al. (2008) for the task of recognising textual entailment, originally proposed by Dagan et al. (2006) in the PASCAL Recognizing Textual Entailment (RTE) Challenge. RTE is a binary classification task consisting in judging whether the meaning of a piece of text (the so-called hypothesis) can be inferred from another piece of text (the entailing text). The original RTE1 Challenge testset consists of 800 text-hypothesis pairs (such as T: *"Chrétien visited Peugeot's newly renovated car factory"*, H: *"Peugeot manufactures cars"*) with a gold standard annotation that classifies each item as either *true* (1)—in case H can be inferred from T—or *false* (0). Exactly 400 items are annotated as 0 and exactly 400 as 1. Bos and Markert (2006) performed an independent expert annotation of this testset, obtaining 95% agreement between the RTE1 gold standard and their own annotation.

The dataset of Snow et al. (2008) includes 10 non-expert annotations for each of the 800 items in the RTE1 testset, collected with Amazon's *Mechanical Turk*. A quick examination of the dataset shows that there are a total of 164 annotators who have annotated between 20 items (124 annotators) and 800 items each (only one annotator). Non-expert annotations with category 1 (rather than 0) are slightly more frequent ($\mathit{Freq}(1) \approx 0.57$).

We have applied our aggregators to this data and compared the outcomes with each other and to the gold standard. The results are summarised in Table 2 and discussed in the sequel. For each pair we report the *observed agreement* A_o (proportion of items on which two annotations agree) and, in brackets, Cohen's *kappa* $\kappa = \frac{A_o - A_e}{1 - A_e}$, with A_e being the expected agreement for independent annotators (Cohen, 1960; Artstein and Poesio, 2008).

Note that there are several variants of the majority rule, depending on how we break ties. In Table 2, $\mathrm{Maj}^{1 \succ 0}$ is the majority rule that chooses 1 in case the number of annotators choosing 1 is equal to the number of annotators choosing 0 (and accordingly for $\mathrm{Maj}^{0 \succ 1}$). For 65 out of the 800 items there has been a tie (i.e., five annotators choose 0 and another five choose 1). This means that the tie-breaking rule used can have a significant impact on results. Snow et al. (2008) work with a majority rule where ties are broken uniformly at random and report an observed agreement (accuracy) between the majority rule and the gold standard of 89.7%. This is confirmed by our results: 89.7% is the mean of 87.5% (our result for $\mathrm{Maj}^{1 \succ 0}$) and 91.9% (our result for $\mathrm{Maj}^{0 \succ 1}$). If we break ties in the optimal way (in view of approximating the gold standard (which of course would not actually be possible without having access to that gold standard), then we obtain an observed agreement of 93.8%, but if we are unlucky and ties happen to get broken in the worst possible way, we obtain an observed agreement of only 85.6%.

For none of our bias-correcting majority rules did we encounter any ties. Hence, for these aggregators the somewhat arbitrary choices we have to make when breaking ties are of no significance, which is an important point in their favour. Observe that all of the bias-correcting majority rules approximate the gold standard better than the majority rule with uniformly random tie-breaking.

[9] $\mathrm{GreedyCR}^t$ is a greedy (rather than optimal) implementation of this basic idea, with the tolerance value t fixing a threshold on (a particular form of) inter-annotator agreement.

[10] Since the annotation task and dataset used for our case study do not involve any interesting integrity constraints, we have not tested any distance-based aggregation rules.

Annotation	Maj$^{1\succ 0}$	Maj$^{0\succ 1}$	ComBCM	DiffBCM	RelBCM	GreedyCR^{0}	GreedyCR^{15}
Gold Standard	87.5% (.75)	91.9% (.84)	91.1% (.80)	91.5% (.81)	90.8% (.80)	86.6% (.73)	92.5% (.85)
Maj$^{1\succ 0}$		91.9% (.84)	88.9% (.76)	94.3% (.87)	94.0% (.87)	87.6% (.75)	91.5% (.83)
Maj$^{0\succ 1}$			96.0% (.91)	97.6% (.95)	96.9% (.93)	89.0% (.78)	96.1% (.92)
ComBCM				94.6% (.86)	94.4% (.86)	88.8% (.75)	93.9% (.86)
DiffBCM					98.8% (.97)	88.6% (.75)	94.8% (.88)
RelBCM						88.4% (.74)	93.8% (.86)
GreedyCR^{0}							90.6% (.81)

Table 2: Observed agreement (and κ) between collective annotations and the gold standard.

Recall that the greedy consensus rule is in fact a family of aggregators: whenever there is more than one item with a maximal majority, we may lock in any one of them. Furthermore, when there is a split majority, then ties may be broken either way. The results reported here refer to an implementation that always chooses the lexicographically first item amongst all those with a maximal majority and that breaks ties in favour of 1. These parameters yield neither the best or the worst approximations of the gold standard. We tested a range of tolerance values. As an example, Table 2 includes results for tolerance values 0 and 15. The coalition found for tolerance 0 consists of 46 annotators who all completely agree with the collective annotation; the coalition found for tolerance 15 consists of 156 annotators who all disagree with the collective annotation on at most 15 items. While GreedyCR^{0} appears to perform rather poorly, GreedyCR^{15} approximates the gold standard particularly well. This is surprising and suggests, on the one hand, that eliminating only the most extreme outlier annotators is a useful strategy, and on the other hand, that a high-quality collective annotation can be obtained from a group of annotators that disagree substantially.[11]

6 Related Work

There is an increasing number of projects using crowdsourcing methods for labelling data. Online *Games with a Purpose*, originally conceived by von Ahn and Dabbish (2004) to annotate images, have been used for a variety of linguistic tasks: Lafourcade (2007) created JeuxDeMots to develop a semantic network by asking players to label words with semantically related words; Phrase Detectives (Chamberlain et al., 2008) has been used to gather annotations on anaphoric coreference; and more recently Basile et al. (2012) have developed the Wordrobe set of games for annotating named entities, word senses, homographs, and pronouns. Similarly, crowdsourcing via microworking sites like Amazon's *Mechanical Turk* has been used in several annotation experiments related to tasks such as affect analysis, event annotation, sense definition and word sense disambiguation (Snow et al., 2008; Rumshisky, 2011; Rumshisky et al., 2012), amongst others.[12]

All these efforts face the problem of how to aggregate the information provided by a group of volunteers into a collective annotation. However, by and large, the emphasis so far has been on issues such as experiment design, data quality, and costs, with little attention being paid to the aggregation methods used, which are typically limited to some form of majority vote (or taking averages if the categories are numeric). In contrast, our focus has been on investigating different aggregation methods for arriving at a collective annotation.

Our work has connections with the literature on inter-annotator agreement. Agreement scores such as *kappa* are used to assess the quality of an annotation but do not play a direct role in constructing one single annotation from the labellings of several coders.[13] The methods we have proposed, in contrast, do precisely that. Still, agreement plays a prominent role in some of these methods. In our discussion of distance-based aggregation, we suggested how agreement can be used to select a subset of annotators whose individual annotations are minimally distant from the resulting collective annotation. Our greedy consensus rule also makes use of agreement to ensure a minimum level of consensus. In both cases, the aggregators have the effect of disregarding some outlier annotators.

[11] Recall that 124 out of 164 coders only annotated 20 items each; a tolerance value of 15 thus is fairly lenient.

[12] See also the papers presented at the NAACL 2010 Workshop on Creating Speech and Language Data with Amazon's *Mechanical Turk* (`tinyurl.com/amtworkshop2010`).

[13] Creating a gold standard often involves adjudication of disagreements by experts, or even the removal of cases with disagreement from the dataset. See, e.g., the papers cited by Beigman Klebanov and Beigman (2009).

Other researchers have explored ways to directly identify "low-quality" annotators. For instance, Snow et al. (2008) and Raykar et al. (2010) propose Bayesian methods for identifying and correcting annotators' biases, while Ipeirotis et al. (2010) propose an algorithm for assigning a quality score to annotators that distinguishes intrinsic error rate from an annotator's bias. In our approach, we do not directly rate annotators or recalibrate their annotations—rather, some outlier annotators get to play a marginal role in the resulting collective annotation as a side effect of the aggregation methods themselves.

Although in our case study we have tested our aggregators by comparing their outcomes to a gold standard, our approach to collective annotation itself does *not* assume that there is in fact a ground truth. Instead, we view collective annotations as reflecting the views of a community of speakers.[14] This contrasts significantly with, for instance, the machine learning literature, where there is a focus on estimating *the hidden true label* from a set of noisy labels using maximum-likelihood estimators (Dawid and Skene, 1979; Smyth et al., 1995; Raykar et al., 2010).

In application domains where it is reasonable to assume the existence of a ground truth and where we are able to model the manner in which individual judgments are being distorted relative to this ground truth, social choice theory provides tools (using again maximum-likelihood estimators) for the design of aggregators that maximise chances of recovering the ground truth for a given model of distortion (Young, 1995; Conitzer and Sandholm, 2005). In recent work, Mao et al. (2013) have discussed the use of these methods in the context of crowdsourcing. Specifically, they have designed an experiment in which the ground truth is defined unambiguously and known to the experiment designer, so as to be able to extract realistic models of distortion from the data collected in a crowdsourcing exercise.

7 Conclusions

We have presented a framework for combining the expertise of speakers taking part in large-scale annotation projects. Such projects are becoming more and more common, due to the availability of online crowdsourcing methods for data annotation. Our work is novel in several respects. We have drawn inspiration from the field of social choice theory to formulate a general formal model for aggregation problems, which we believe sheds light on the kind of issues that arise when trying to build annotated linguistic resources from a potentially large group of annotators; and we have proposed several families of concrete methods for aggregating individual annotations that are more fine-grained that the standard majority rule that so far has been used across the board. We have tested some of our methods on a gold standard testset for the task of recognising textual entailment.

Our aim has been conceptual, namely to point out that it is important for computational linguists to reflect on the methods used when aggregating annotation information. We believe that social choice theory offers an appropriate general methodology for supporting this reflection. Importantly, this does not mean that the concrete aggregation methods developed in social choice theory are immediately applicable or that all the axioms typically studied in social choice theory are necessarily relevant to aggregating linguistic annotations. Rather, what we claim is that it is the *methodology* of social choice theory which is useful: to formally state desirable properties of aggregators as axioms and then to investigate which specific aggregators satisfy them. To put it differently: at the moment, researchers in computational linguistics simply use some given aggregation methods (almost always the majority rule) and judge their quality on how they fare in specific experiments—but there is no principled reflection on the methods themselves. We believe that this should change and hope that the framework outlined here can provide a suitable starting point.

In future work, the framework we have presented here should be tested more extensively, not only against a gold standard but also in terms of the usefulness of the derived collective annotations for training supervised learning systems. On the theoretial side, it would be interesting to study the axiomatic properties of the methods of aggregation we have proposed here in more depth and to define axiomatic properties of aggregators that are specifically tailored to the task of collective annotation of linguistic resources.

[14] In some domains, such as medical diagnosis, it makes perfect sense to assume that there is a ground truth. However, in tasks related to linguistic knowledge and language use such an assumption seems far less justified. Hence, a collective annotation may be the closest we can get to a representation of the linguistic knowledge/use of a linguistic community.

References

James Allen and Mark Core, 1997. *DAMSL: Dialogue Act Markup in Several Layers*. Discourse Resource Initiative.

Kenneth J. Arrow, Armatya K. Sen, and Kotaro Suzumura, editors. 2002. *Handbook of Social Choice and Welfare*. North-Holland.

Kenneth J. Arrow. 1963. *Social Choice and Individual Values*. John Wiley and Sons, 2nd edition. First edition published in 1951.

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

Michel Balinski and Rida Laraki. 2011. *Majority Judgment: Measuring, Ranking, and Electing*. MIT Press.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. A platform for collaborative semantic annotation. In *Proc. 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, pages 92–96.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Johan Bos and Katja Markert. 2006. Recognising textual entailment with robust logical inference. In *Machine Learning Challenges*, volume 3944 of *LNCS*, pages 404–426. Springer-Verlag.

Steven J. Brams and Peter C. Fishburn. 2002. Voting procedures. In Kenneth J. Arrow, Armartya K. Sen, and Kotaro Suzumura, editors, *Handbook of Social Choice and Welfare*. North-Holland.

Jean Carletta, Stephen Isard, Anne H. Anderson, Gwyneth Doherty-Sneddon, Amy Isard, and Jacqueline C. Kowtko. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Addressing the resource bottleneck to create large-scale annotated texts. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 375–380. College Publications.

Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. 2008. Preference handling in combinatorial domains: From AI to social choice. *AI Magazine*, 29(4):37–46.

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

Vincent Conitzer and Tuomas Sandholm. 2005. Common voting rules as maximum likelihood estimators. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*.

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 Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer errorrates using the EM algorithm. *Applied Statistics*, 28(1):20–28.

Elad Dokow and Ron Holzman. 2010. Aggregation of binary evaluations. *Journal of Economic Theory*, 145(2):495–511.

Ulle Endriss, Maria Silvia Pini, Francesca Rossi, and K. Brent Venable. 2009. Preference aggregation over restricted ballot languages: Sincerity and strategy-proofness. In *Proc. 21st International Joint Conference on Artificial Intelligence (IJCAI-2009)*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proc. 47th Annual Meeting of the Association for Computational Linguistics (ACL-2009)*, pages 10–18.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Gerald J. Glasser. 1959. Game theory and cumulative voting for corporate directors. *Management Science*, 5(2):151–156.

Umberto Grandi and Ulle Endriss. 2011. Binary aggregation with integrity constraints. In *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*.

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

Mukund Jha, Jacob Andreas, Kapil Thadani, Sara Rosenthal, and Kathleen McKeown. 2010. Corpus creation for new genres: A crowdsourced approach to PP attachment. In *Proc. NAACL-HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 13–20.

John Kemeny. 1959. Mathematics without numbers. *Daedalus*, 88:577–591.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on senseval. *Computers and the Humanities*, 34(1):1–13.

Sébastien Konieczny and Ramón Pino Pérez. 2002. Merging information under constraints: A logical framework. *Journal of Logic and Computation*, 12(5):773–808.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition with the JeuxDeMots prototype. In *Proc. 7th International Symposium on Natural Language Processing*.

Christian List and Philip Pettit. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18(1):89–110.

Andrew Mao, Ariel D. Procaccia, and Yiling Chen. 2013. Better human computation through principled voting. In *Proc. 27th AAAI Conference on Artificial Intelligence*.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Michael K. Miller and Daniel Osherson. 2009. Methods for distance-based judgment aggregation. *Social Choice and Welfare*, 32(4):575–601.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Maria Silvia Pini, Francesca Rossi, K. Brent Venable, and Toby Walsh. 2009. Aggregating partially ordered preferences. *Journal of Logic and Computation*, 19(3):475–502.

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

Sara Rosenthal, William Lipovsky, Kathleen McKeown, Kapil Thadani, and Jacob Andreas. 2010. Towards semi-automated annotation for prepositional phrase attachment. In *Proc. 7th International Conference on Language Resources and Evaluation (LREC-2010)*.

Anna Rumshisky, Nick Botchan, Sophie Kushkuley, and James Pustejovsky. 2012. Word sense inventories by non-experts. In *Proc. 8th International Conference on Language Resources and Evaluation (LREC-2012)*.

Anna Rumshisky. 2011. Crowdsourcing word sense definition. In *Proc. ACL-HLT 5th Linguistic Annotation Workshop (LAW-V)*.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1995. Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems*, pages 1085–1092.

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.

Alan D. Taylor. 2005. *Social Choice and the Mathematics of Manipulation*. Cambridge University Press.

T. Nicolaus Tideman. 1987. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4(3):185–206.

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

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326. ACM.

H. Peyton Young. 1995. Optimal voting rules. *Journal of Economic Perspectives*, 9(1):51–64.