Collective Annotation: From Crowdsourcing to Social Choice

Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam

[joint work with Raquel Fernández, Justin Kruger and Ciyang Qing]

Outline

Ideas from *social choice theory* can be used for *collective annotation* of data obtained by means *crowdsourcing*.

- Annotation and Crowdsourcing (in Linguistics and other fields)
- Formal Framework: Axiomatics of Collective Annotation
- Three Concrete Methods of Aggregation
- Results from Three Case Studies in Linguistics

The talk is based on the three papers cited below.

U. Endriss and R. Fernández. Collective Annotation of Linguistic Resources: Basic Principles and a Formal Model. Proc. ACL-2013.

J. Kruger, U. Endriss, R. Fernández, and C. Qing. Axiomatic Analysis of Aggregation Methods for Collective Annotation. Proc. AAMAS-2014.

C. Qing, U. Endriss, R. Fernández, and J. Kruger. Empirical Analysis of Aggregation Methods for Collective Annotation. Proc. COLING-2014.

Annotation and Crowdsourcing

Disciplines such as computer vision and computational linguistics require large corpora of annotated data.

Examples from linguistics: grammaticality, word senses, speech acts

People need corpora with *gold standard* annotations:

- set of *items* (e.g., text fragment with one utterance highlighted)
- assignment of a *category* to each item (e.g., it's a *question*)

Classical approach: ask a handful of experts (who hopefully agree).

Modern approach is to use *crowdsourcing* (e.g., Mechanical Turk) to collect annotations: fast, cheap, more judgments from more speakers.

<u>But:</u> how to *aggregate* individual annotations into a gold standard?

- some work using machine learning approaches
- dominant approach: for each item, adopt the *majority* choice

Social Choice Theory

Aggregating information from individuals is what *social choice theory* is all about. Example: aggregation of preferences in an election.

F: vector of individual preferences $\mapsto$ election winner
F: vector of individual annotations $\mapsto$ *collective annotation*

Research agenda:

- develop a variety of *aggregation methods* for collective annotation
- *analyse* those methods in a principled manner, as in SCT
- understand features specific to applications via *empirical studies*

Formal Model

An annotation task has three components:

- infinite set of *agents* N
- finite set of *items* J
- finite set of *categories* K

A finite subset of agents annotate some of the items with categories (one each), resulting is a *group annotation* $A \subseteq N \times J \times K$.

$(i, j, k) \in A$ means that agent i annotates item j with category k.

An *aggregator* F is a mapping from group annotations to annotations:

$$F : 2^{N \times J \times K}_{<\omega} \to 2^{J \times K}$$

Axioms

In social choice theory, an *axiom* is a formal rendering of an intuitively desirable property of an aggregator F. Examples:

- *Nontriviality:* $|A \restriction j| > 0$ should imply $|F(A) \restriction j| > 0$
- *Groundedness:* $\mathrm{cat}(F(A) \restriction j)$ should be a subset of $\mathrm{cat}(A \restriction j)$
- *Item-Independence:* $F(A) \restriction j$ should be equal to $F(A \restriction j)$
- *Agent-Symmetry:* $F(\sigma(A)) = F(A)$ for all $\sigma : N \to N$
- *Category-Symmetry:* $F(\sigma(A)) = \sigma(F(A))$ for all $\sigma : K \to K$
- *Positive Responsiveness:* $k \in \mathrm{cat}(F(A) \restriction j)$ and $(i, j, k) \notin A$ should imply $\mathrm{cat}(F(A \cup (i, j, k)) \restriction j) = \{k\}$

Reminder: annotation A, agents $i \in N$, items $j \in J$, categories $k \in K$

Characterisation Results

An elegant characterisation of the most basic aggregation rule:

Theorem 1 (Simple Plurality) *An aggregator F is nontrivial, item-independent, agent-symmetric, category-symmetric, and positively responsive iff F is the simple plurality rule:*

$$F : A \mapsto \{(j, k^\star) \in J \times K \mid k^\star \in \operatorname*{argmax}_{k \in \mathrm{cat}(A\restriction j)} |A \restriction j, k|\}$$

An argument for describing rules in terms of weights:

Theorem 2 (Weights) *An aggregator F is nontrivial and grounded iff it is a weighted rule (fully defined in terms of weights $w_{i,j,k}$).*

Concrete Aggregation Rules

We have three proposals for concrete aggregation rules that are more sophisticated than the simple plurality rule and that try to account for the *reliability of individual annotators* in different ways:

- Bias-Correcting Rules
- Greedy Consensus Rules
- Agreement-Based Rule

Proposal 1: Bias-Correcting Rules

If an annotator appears to be *biased* towards a particular category, then we could try to correct for this bias during aggregation.

- $\mathrm{Freq}_i(k)$: relative frequency of annotator i choosing category k
- $\mathrm{Freq}(k)$: relative frequency of k across the full profile

$\mathrm{Freq}_i(k) > \mathrm{Freq}(k)$ suggests that i is biased towards category k.

A *bias-correcting rule* tries to account for this by varying the weight given to k-annotations provided by annotator i:

- **Diff** (difference-based): $w_{ik} = 1 + \mathrm{Freq}(k) - \mathrm{Freq}_i(k)$
- **Rat** (ratio-based): $w_{ik} = \mathrm{Freq}(k) \,/\, \mathrm{Freq}_i(k)$
- **Com** (complement-based): $w_{ik} = 1 + 1 \,/\, |K| - \mathrm{Freq}_i(k)$
- **Inv** (inverse-based): $w_{ik} = 1 \,/\, \mathrm{Freq}_i(k)$

For comparison: the *simple majority rule* SPR always assigns weight 1.

Proposal 2: Greedy Consensus Rules

If there is *(near-)consensus* on an item, we should adopt that choice. And: we might want to classify annotators who disagree as *unreliable*.

The *greedy consensus rule* **GreedyCR**^{t} (with *tolerance threshold* t) repeats two steps until all items are decided:

(1) *Lock in* the majority decision for the item with the strongest majority not yet locked in.

(2) *Eliminate* any annotator who disagrees with more than t decisions.

Variations are possible: any nondecreasing function from disagreements with locked-in decisions to annotator weight might be of interest.

Greedy consensus rules appar to be good at recognising *item difficulty*.

Proposal 3: Agreement-Based Rule

Suppose each item has a *true* category (its *gold standard*). If we knew it, we could compute each annotator i's *accuracy* acc_i.

If we knew acc_i, we could compute annotator i's *optimal weight* w_i (using maximum likelihood estimation, under certain assumptions):

$$w_i \quad = \quad \log \frac{(|K|-1) \cdot \mathrm{acc}_i}{1 - \mathrm{acc}_i}$$

But we don't know acc_i. However, we can try to *estimate* it as annotator i's *agreement* agr_i with the plurality outcome:

$$\mathrm{agr}_i \quad = \quad \frac{|\{j \in J \mid i \text{ agrees with SPR on } j\}| + 0.5}{|\{j \in J \mid i \text{ annotates } j\}| + 1}$$

The agreement rule **Agr** thus uses weights $w'_i = \log \frac{(|K|-1) \cdot \mathrm{agr}_i}{1 - \mathrm{agr}_i}$.

Empirical Analysis

We have implemented our three types of aggregation rules and compared the results they produce to *existing gold standard* annotations for three tasks in computational linguistics:

- RTE: *recognising textual entailment* (2 categories)
- PSD: *proposition sense disambiguation* (3 categories)
- QDA: *question dialogue acts* (4 categories)

For RTE we used readily available crowdsourced annotations.
For PSD and QDA we collected new crowdsourced datasets.

GreedyCR so far has only been implemented for the binary case.

Case Study 1: Recognising Textual Entailment

In RTE tasks you try to develop algorithms to decide whether a given piece of text entails a given hypothesis. Examples:

Text	Hypothesis	GS
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.	Yahoo bought Overture.	1
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology.	Israel was established in May 1971.	0

We used a dataset collected by Snow et al. (2008):

- Gold standard: 800 items (T-H pairs) with an 'expert' annotation
- Crowdsourced data: 10 AMT annotations per item (164 people)

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. Proc. EMNLP-2008.

Example

An example where GreedyCR^{15} correctly overturns a 7-3 majority against the gold standard (0, i.e., T does *not* entail H):

T:	The debacle marked a new low in the erosion of the SPD's popularity, which began after Mr. Schröder's election in 1998.
H:	The SPD's popularity is growing.

The item ends up being the 631st to be considered:

Annotator	Choice	disagr's	In/Out
AXBQF8RALCIGV	1	83	×
A14JQX7IFAICP0	1	34	×
A1Q4VUJBMY78YR	1	81	×
A18941IO2ZZWW6	1	148	×
AEX5NCH03LWSG	1	19	×
A3JEUXPU5NEHXR	**0**	2	✓
A11GX90QFWDLMM	1	143	×
A14WWG6NKBDWGP	**1**	1	✓
A2CJUR18C55EF4	**0**	2	✓
AKTL5L2PJ2XCH	**0**	1	✓

Case Study 2: Preposition Sense Disambiguation

The PSD task is about choosing the sense of the preposition "*among*" in a given sentence, out of three possible senses from the ODE:

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

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

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

We crowdsourced data for a corpus with an existing GS annotation:

- Gold standard: 150 items (sentences) from *SemEval 2007*
- Crowdsourced data: 10 AMT annotations per item (45 people)

K.C. Litkowski and O. Hargraves. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. Proc. SemEval-2007.

Case Study 3: Question Dialogue Acts

The QDA task consists in selecting a *question dialogue act*, for a highlighted utterance in a dialogue fragment, out of four possibilities:

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

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

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

(4) **Rhetorical:** Questions that do not need to be answered, but are asked only to make a point, e.g., *"If I had a pet, how could I work?"*

We crowdsourced data for a corpus with an existing GS annotation:

- Gold standard: 300 questions from the *Switchboard Corpus*
- Crowdsourced data: 10 AMT annotations per item (63 people)

D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL: Shallow-Discourse-Function-Annotation Coders Manual. Univ. of Colorado Boulder, 1997.

Case Studies: Results

How well did we do? Observed *agreement* with the gold standard annotation (any ties are counted as instances of disagreement):

- Recognising Textual Entailment (two categories):
  - SPR: 85.6%
  - Best BCR's: Com 91.6%, Diff 91.5%
  - Agr: 93.3%
  - GreedyCR^0: 86.6%, GreedyCR^{15}: 92.5%
- Preposition Sense Disambiguation (three categories):
  - SPR: 81.3% [caveat: gold standard appears to have errors]
  - Best BCR: Rat 84%, Diff 83.3%
  - Agr: 82.7%
- Question Dialogue Acts (four categories):
  - SPR: 85.7%
  - Best BCR: Inv 87.7% [shared bias $\rightsquigarrow$ agent-indep. rules better]
  - Agr: 86.7%

Last Slide

We took inspiration from *social choice theory* to formulate a model for aggregating expertise of speakers in *annotation projects*. Specifically:

- Provided *axiomatic characterisation* of simple plurality rule and of family of all rules that can be described via weights.
- Proposed three families of *aggregation methods* that are more sophisticated than the standard majority rule, by accounting for the *reliability of individual annotators*.
- Empirical results show small but *robust improvements* over the simple plurality/majority rule (also requiring *fewer annotators*).

Papers and *crowdsourced data* are available here:

`http://www.illc.uva.nl/Resources/CollectiveAnnotation/`