

# Collective Annotation: Applying Voting Theory to Computational Linguistics

Ulle Endriss

Institute for Logic, Language and Computation

University of Amsterdam

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

## Students Involved

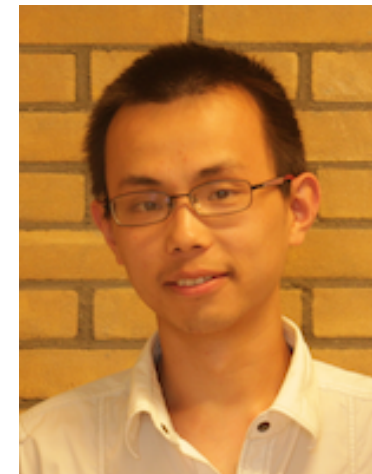


### **Justin Kruger (Master of Logic 2014)**

- ▶ Bachelor Philosophy, University of St Andrews, 2011
- ▶ Now: PhD Computer Science and Decision Analysis, Paris-Dauphine University

### **Ciyang Qing (Master of Logic 2014)**

- ▶ Bachelor Computer Science, Peking University, 2012
- ▶ Now: PhD Linguistics & Cognitive Science, Stanford University



## Challenge: Annotation for Linguistics

Imagine a researcher in computational linguistics, working on designing a new voice-controlled personal assistant, wants to understand what distinguishes *rhetorical questions* from other kinds of questions . . .

They will need a lot of *annotated data*, like this:

B: [Noise] Yeah.

B: It, it's one of those necessities of life that we all have to, you know, pay taxes but, although it is kind of a pain sometimes though.

A: It's just scary though about, you know. —

**A: How high are the taxes going to be when my children are my age?**

B: Uh-huh.

A: You know, that, that's, that's scary too.

Yes-No  Wh  Declarative  Rhetorical

# Collecting Raw Annotations: Crowdsourcing

The screenshot shows the Amazon Mechanical Turk interface. At the top left is the Amazon Mechanical Turk logo with the text 'Artificial Intelligence'. Navigation buttons include 'Your Account', 'HITS', and 'Qualifications'. A status bar indicates '244,501 HITS available now'. A search bar contains 'Find HITS containing' followed by a text input field and a 'GO' button. Filter options include 'for which you are qualified' and 'require Master Qualification'. A timer shows '00:00:00 of 15 minutes'. Action buttons for 'Accept HIT' and 'Skip HIT' are present. Summary statistics show 'Total Earned: Unavailable' and 'Total HITS Submitted: 0'.

**1. Yes-No Questions** [Show examples](#)

Questions that *have the standard form* of a question and that *could be* answered by saying "yes" or "no" (Careful! They are *not always answered in this way*. It only matters whether they could).

**2. Wh Questions** [Show examples](#)

Questions that *have the standard form* of a question and that ask for *specific* information by means of a question word such as "what", "who", "which", "when", "where" or "how".

**3. Declarative Questions** [Show examples](#)

Questions that *don't have the standard form* of a question (they look more like statements) but *nevertheless ask for some answer*, which could be a "yes"/"no" answer or more specific information.

**4. Rhetorical Questions** [Show examples](#)

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* (often negative), for the sake of encouraging the listener to consider an issue.

**In this task you are asked to classify the questions in 10 fragments of dialogues, according to the definitions on the left (with examples):**

**Read the definitions of different types of questions on the left carefully, as well as the examples that follow. Please choose the type that is closest to the usage of the question marked in bold in each dialogue fragment below. (You should always classify what is marked in bold, even if sometimes it is without a question mark!)**

Dialogue 1.  
 A: and the other one doesn't.  
 A: And you're right, they do get bored, uh, really fast, if they already know what you're talking about.  
 A: What do you propose that they do?  
**A: What, what is your suggestions?**  
 B: The educators need to be a little bit more open minded as well as innovative in dealing with, uh, the various students to get the maximum potential out of the person.  
 A: Uh-huh.  
 A: Out of each child.

Yes-No  Wh  Declarative  Rhetorical

## Idea: Collective Annotation as Social Choice

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

$F$ : vector of individual preferences  $\mapsto$  election winner

$F$ : vector of individual annotations  $\mapsto$  collective annotation

## Example: Estimating Accuracy as Agreement

Naïve approach: *majority voting*. We have developed several more sophisticated aggregation rules. Here is one:

- (1) Assume *annotator*  $i$  makes correct choice with *probability*  $p_i$ , and each of the wrong choices with equal probability  $(1 - p_i)/(k - 1)$ .
- (2) Use *weighted majority voting*, giving more weight to annotators  $i$  with higher accuracy  $p_i$ . How much more? *Maximum likelihood* for:

$$weight_i = \log \frac{(k - 1) \cdot p_i}{1 - p_i}$$

Great ... except that actually *we don't know* any of the  $p_i$ 's!

- (3) But we can try to *estimate* the *accuracy*  $p_i$  of annotator  $i$  as her observed *agreement with the simple majority rule*:

$$p_i \approx \frac{\# \text{ items where } i \text{ and majority rule agree} + 0.5}{\# \text{ items annotated by } i + 1}$$

## Results

*Majority voting* with *10 annotations* per item achieves *85% accuracy*, relative to an existing corpus annotated manually by experts.

*Our rule* achieves the *same accuracy* with just *6 annotations* per item.

For more rules, results, our papers, and our crowdsourced data, see:

<http://www.illc.uva.nl/Resources/CollectiveAnnotation/>

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.