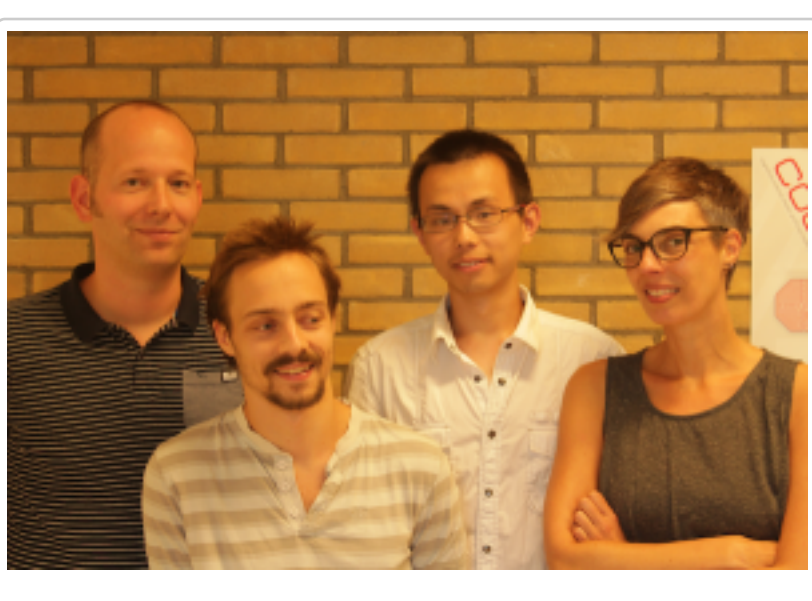
COLING Blog Series: #1 Collective Annotation – Unexpected Connections between Computational Linguistics and Voting Theory

11/08/2014 Louise Irwin COLING2014, Intelligent Content



Ulle Endriss, Justin Kruger, Ciyang Qing, and Raquel Fernández, Institute for Logic, Language and Computation, University of Amsterdam

In anticipation of COLING 2014, which is just weeks away, the CNGL Centre for Global Intelligent Content is publishing a special COLING guest blog series to introduce paper authors and their research.

Winning papers at each COLING conference are selected by a committee, consisting of the Scientific Advisory Board and Program Chairs. This year, IBM Watson, an exciting new R&D division, based in Dublin, and dedicated to developing cloud-based cognitive applications and services in both research and industry communities, is supporting the Best Paper Awards.

To start the ball rolling, the first post in this series is by Raquel Fernández and Ulle Endriss. Raquel and Ulle are senior scientists at the Institute for Logic, Language and Computation (ILLC), University of Amsterdam, and they introduce their research entitled 'Empirical Analysis of Aggregation Methods for Collective Annotation' by Ciyang Qing, Ulle Endriss, Raquel Fernández, and Justin Kruger, which received a Honourable Mention Award.

Collective Annotation: Unexpected Connections between Computational Linguistics and Voting Theory

It is a bit of a cliché, but the Internet has changed our lives in all sorts of ways. One of them is how scientists collect data. Imagine a linguist seeking to understand how people make judgments regarding their native language: for example, when does someone judge a given question to be a rhetorical question rather than a genuine one? In the past, our linguist would have had to run a small-scale experiment, with the participants most likely being students drawn from her Linguistics 101 class. Today, she can instead reach speakers all over the globe and from all walks of life, collecting lots of data on people's linguistic judgments over the Internet.

But what do you then do with all this new data? It might be rather noisy: the participants will not have been trained in linguistics, they may not have fully grasped the instructions given, and some may in fact not care very much about the advancement of science and be content with collecting their participation fee as quickly as possible. But we would still like to arrive at a high-quality judgment about how the population of native speakers of a given language categorise a given item, even if the individual speakers do not all agree with each other.

Our research deals with this problem. Suppose you have collected the judgments of a large number of people regarding a large number of different items. Each item needs to be assigned a category — e.g., "rhetorical" or "genuine", if the items are questions. For example, in one

of our experiments a total of 63 participants submitted judgments regarding 300 items, with each item getting annotated by 10 individuals and each individual annotating between 10 and 200 items.

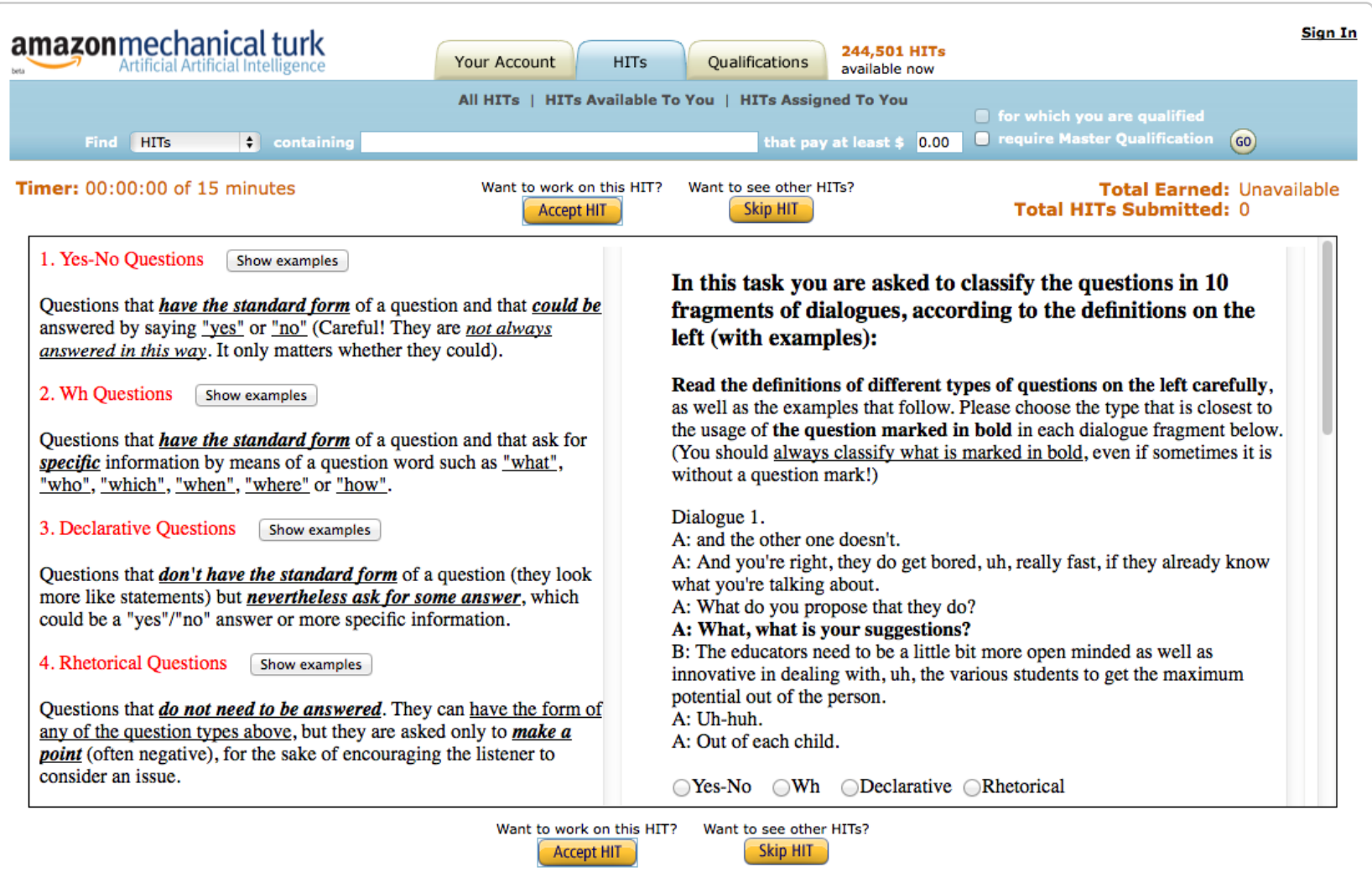
The question then is: how do you aggregate this data to arrive at a single category for each item?

We have approached this question as a problem of voting. Think of it as an election, with people voting for their favourite judgment for every given item. The properties of different voting rules are studied in a field known as social choice theory. While no existing voting rule intended for political elections directly fits our problem, the general principles developed in social choice theory can be applied also here.

> These principles have helped us to design novel methods for aggregating individual judgments that produce better results than the naïve method of simply choosing for every item the category selected most often by the participants.

The paper presented at the COLING conference in Dublin this August is one of a series of papers developing these ideas. You can find out more about this ongoing project on the Collective Annotation website.

To date we have applied this new methodology for collectively annotating data only to problems in the domain of linguistics. But the methodology itself is much more general than that, and we believe that it has a lot of potential also for other fields where multiple diverse judgements may play a role, such as image recognition or medical diagnosis. We would be delighted to hear from anyone who may have suggestions in this respect. Our contact details are available from our homepages listed below.



Crowdsourcing interface for collecting speaker judgments

About the authors

Raquel Fernández and Ulle Endriss are senior scientists at the Institute for Logic, Language and Computation (ILLC) of the University of Amsterdam. Raquel has a background in computational linguistics and cognitive science, and works on topics concerned with semantics, dialogue, and language coordination. Ulle conducts research in logic and artificial intelligence, focussing on questions related to economics and political science.

Ciyang Qing and Justin Kruger are Master of Logic students at the ILLC. In September 2014, Ciyang will start a PhD in Linguistics at Stanford University and Justin a PhD in Computer Science at Paris-Dauphine University.

Share this post: on Twitter on Facebook on Google+

Ciyang Qing | CNGL | COLING | COLING 2014 | Coling Honourable Mention | Collective Annotation | Computational Linguistics | DCU | Dublin City University | ILLC | Institute for Logic | Justin Kruger | Language and Computation | Raquel Fernández | The CNGL Centre for Global Intelligent Content | Ulle Endriss | University of Amsterdam | voting theory

LEAVE A REPLY

Your email address will not be published. Required fields are marked *

Name *

Email *

Website

Comment

You may use these HTML tags and attributes: `<a href="" title=""> <abbr title=""> <acronym title=""> <b> <blockquote cite=""> <cite> <code> <del datetime=""> <em> <i> <q cite=""> <strike> <strong>`

POST COMMENT

‹ CNGL Interview: Teresa Lynn – Irish Language for Social Media and Machine Translation

COLING Blog Series: #2 Learning to Distinguish Hypernyms and Co-Hyponyms ›

Search...

SUBSCRIBE TO THE CNGL BLOG

Your Email

SEND

CATEGORIES

RECENT POSTS

ARCHIVES