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

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

[joint work with Raquel Fernández]

Outline

- Annotation and Crowdsourcing in Linguistics
- Proposal: Use Social Choice Theory
- Two New Methods of Aggregation
- Results from a Case Study on Textual Entailment

Annotation and Crowdsourcing in Linguistics

To test theories in linguistics and to benchmark algorithms in NLP, we require information on the *linguistic judgments of speakers*. Examples: 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 an agreement act)

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

But: how to *aggregate* individual annotations into a gold standard?

- some work on maximum likelihood estimators
- 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.

$$F\text{: vector of individual preferences} \mapsto \text{election winner}$$
$$F\text{: vector of individual annotations} \mapsto \textit{collective annotation}$$

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 linguistics via *empirical studies*

For this talk: assume there are just *two categories* (0 and 1).

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.

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

$\text{Freq}_i(k) > \text{Freq}(k)$ 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:

- difference-based: $1 + \text{Freq}(k) - \text{Freq}_i(k)$
- ratio-based: $\text{Freq}(k) \,/\, \text{Freq}_i(k)$

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

Ongoing work: axiomatise this class of rules à la SCT

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.

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

Ongoing work: try to better understand this phenomenon

Case Study: 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 MTurk 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.

Case Study: Results

How did we do? Observed *agreement* with the gold standard:

- Simple Majority Rule (produced 65 ties for 800 items):
  - 89.7% under uniform tie-breaking
  - 85.6% if ties are counted as misses
- Bias-Correcting Rules (no ties encountered):
  - 91.5% for the difference-based rule
  - 90.8% for the ratio-based rule
- Greedy Consensus Rules (for certain implementation choices):
  - 86.6% for tolerance threshold 0 (found coalition of 46/164)
  - 92.5% for tolerance threshold 15 (found coalition of 156/164)

Ongoing work: understand better what performance depends on

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	✓

Last Slide

- Took inspiration from *social choice theory* to formulate model for aggregating expertise of speakers in *annotation projects*.
- Proposed two families of *aggregation methods* that are more sophisticated than the standard majority rule, by accounting for the *reliability of individual annotators*.
- Our broader aim is to reflect on the methods used to aggregate annotation information: social choice theory can help.