Collective Annotation of Linguistic Resources: Basic Principles and a Formal Model [1]

Ulle Endriss Raquel Fernández

Institute for Logic, Language and Computation (ILLC)
University of Amsterdam

Abstract

Crowdsourcing provides new ways of cheaply and quickly gathering large amounts of information contributed by volunteers online. This method has revolutionised the collection of labelled data, in computational linguistics and elsewhere. However, to create annotated linguistic resources from crowdsourced data we face the challenge of having to combine the judgements of a potentially large group of annotators. Here we put forward the idea of using principles of social choice theory to design new methods for aggregating linguistic annotations provided by individuals into a single collective annotation.

1 Social Choice Theory for Computational Linguistics

Research in computational linguistics relies on the availability of linguistic resources in the form of annotated corpora, to be able to test linguistic theories and to evaluate the performance of algorithms for natural language processing tasks. 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 this "wisdom of the crowds" can be used to develop useful annotated linguistic resources. Indeed, those who have investigated this increasingly important issue have only used very simple methods based on majority voting to combine the judgments of individual annotators.

The problem of deriving a single *collective annotation* from a set of diverse judgments is an aggregation problem that shares certain similarities with problems studied in *social choice theory* [1], a theoretical framework for amalgamating the preferences of several individuals into a collective decision (e.g., in a political election). We propose to exploit this similarity and to use social choice theory as a methodological framework for a systematic reflection on the methods used to aggregate annotation information. One advantage of this approach is that it would allow us to make desirable properties of such methods more explicit.

As a first step towards this general goal, we have defined several new methods of aggregation and conducted a case study testing their performance on an existing dataset for textual entailment.

2 Aggregation Methods for Annotators of Varying Reliability

An annotation task consists of a set of *items*, each of which is associated with a set of possible *categories* [2]. For simplicity, suppose there are just two categories, 0 and 1. Each *annotator* is presented with a subset of the items and asked to label each of them with either 0 or 1. An *aggregator* is a function that maps any such

[1] The original paper has been published in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 539–549, 2013.

profile into a single *collective annotation* of the full set of items. An example is the *simple majority rule:* label a given item with 0 if more annotators chose 0 than chose 1; label it with 1 if the opposite is the case; and toss a coin in case both categories occur equally often. Here we sketch two slightly more sophisticated aggregators that try to account for the fact that not all annotators will be equally reliable.

Bias-Correcting Majority Rules. If someone annotates most items with 0, then we might want to assign less significance to that choice for any given item. That is, if an annotator appears to be *biased* towards a particular category, then we could try to correct for this bias during aggregation by adapting her weight.

For a given profile of annotations, let $\mathit{Freq}_i(X)$ be the relative frequency of annotator i choosing category $X \in \{0, 1\}$ (e.g., if i annotated 100 items and on 38 occasions chose category 1, then $\mathit{Freq}_i(1) = 0.38$). Similarly, let $\mathit{Freq}(X)$ denote the frequency of X across the entire profile. For example, if $\mathit{Freq}_i(1) > \mathit{Freq}(1)$, then this suggests that annotator i is biased towards using category 1. We define the *difference-based bias-correcting majority rule* as the weighted majority rule where the vote of annotator i in favour of category X is weighted by $1 + \mathit{Freq}(X) - \mathit{Freq}_i(X)$. (Other options are possible as well, e.g., to use the ratio $\mathit{Freq}(X)/\mathit{Freq}_i(X)$ as the weight instead.)

Greedy Consensus Rules. If for a given item there is almost complete consensus amongst those annotators that annotated it, then we should probably adopt their choice for the collective annotation. Furthermore, the fact that there is almost complete consensus for one item may cast doubts on the reliability of annotators 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 annotations of other items are concerned. In practice, completely eliminating these annotators might be too radical an approach. However, we might want to exclude an annotator from further consideration when her annotation has been different from the majority view at least t times (for some *tolerance value* t).

Our *greedy consensus rules* implement this basic idea by iterating the following two steps: (1) find the item-category pair with the strongest majority and lock in that pair for the collective annotation; (2) eliminate all annotators from the profile that disagree with the pairs locked in so far on more than t items.

3 A Case Study: Recognising Textual Entailment

We have carried out a case study using a dataset created by Snow et al. [3] for the task of *recognising textual entailment*. It 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 (produced by human experts) that classifies each item as either *true* (1)—in case H can be inferred from T—or *false* (0). The dataset includes 10 annotations per item, sourced from a total of 164 non-expert annotators through Amazon's *Mechanical Turk* (each of the 164 individuals annotated between 20 and 800 items).

The simple majority rule (with random tie-breaking) agrees with the gold standard on 89.7% of the items. Our difference-based bias-correcting majority rule achieves an observed agreement of 91.5%. For a (fairly lenient) tolerance value of $t = 15$, our greedy consensus rule achieves 92.5%. These are good results that suggest that these and similar aggregation rules deserve further attention.

References

[1] K.J. Arrow, A.K. Sen, and K. Suzumura, editors. *Handbook of Social Choice and Welfare*. North-Holland, 2002.

[2] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[3] 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. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 254–263, 2008.