



Crowd Sourced Word Sense Annotation

How can we annotate a language database of millions of words? Crowd sourcing techniques may be the answer.

The MASC corpus (<http://www.anc.org/MASC>) is a database of half a million words drawn from written and spoken data sources. MASC is a completely open resource with many genres and modalities of language data, and many types of annotation. For its word sense sentence corpus, trained annotators used WordNet senses to annotate approximately 1000 instances each of over 100 words. In this word inventory representing nouns, verbs and adjectives, the average number of senses per word is 7.2.

At Columbia University, we have been using the Amazon Mechanical Turk system (a framework for crowd source collaboration – see <http://aws.amazon.com/mturk>) to collect the same kind of word-sense annotations on a subset of MASC for a pilot study of half the MASC words. This allows us to conduct a deeper investigation into word sense annotation.

Our previous work showed that ability of annotators to achieve high interannotator reliability varied from word to word, but was independent of the number of senses. In our current work, we apply Bayesian methods to assess the accuracy of individual annotators, and to infer a true label for each instance from a set of labels on the instance from many annotators. A set of labels per instance, each label from a different annotator, provides a much more nuanced representation of variation across words, and across instances for a given word. This work has been done jointly with Bob Carpenter from the Columbia University Department of Statistics.

Rebecca Passonneau is a senior research scientist in the Center for Computational Learning Systems at Columbia University. She has a Ph.D. in Linguistics from University of Chicago, and her main research interests are in computational linguistics and natural language processing.

| | |
|---------------|--|
| Date: | Thursday, March 21, 2013, 8:00 pm. (Refreshments and networking at 7:30 pm.) |
| Place: | Small Auditorium, Room CS 105 Computer Science Building, Princeton University |
| Information: | Dennis Mancl (908) 582-7086, Jan Buzydlowski (610) 902-8343 |
| On-line info: | http://PrincetonACM.acm.org |

All Princeton ACM / IEEE-CS meetings are open to the public. Students and their parents are welcome. There is no admission charge, and refreshments are served.

A pre-meeting dinner is held at 6:00 p.m. at Ruby Tuesday's Restaurant on Route 1. Please send email to princetonacm@acm.org in advance if you plan to attend the dinner.

