Categorizing Modals with Amazon Mechanical Turk

Simonson, Dan; Rubinstein, Aynat; Chung, Joo; Harner, Hillary; Katz, E. Graham; Portner, Paul

Georgetown University

des62@georgetown.edu

1. Introduction

Since the summer of 2010, our research group in the semantics of gradable modal expressions has been hiring workers on Amazon Mechanical Turk to categorize modal expressions. We were motivated with two goals in mind: as part of a larger project involving gradable modals, we wanted to see if turkers—non-linguists—could at least perform the simple task of categorizing modals, and additionally, such information would be a useful addition to any corpus involving modality, like the one we are building of gradable modals.

To evaluate the effectiveness of the turkers, we used Fleiss' kappa scores, a measure of inter-annotator agreement used throughout much of the literature [1]. We also constructed a confusion matrix of the results to look for systematic wanderings from a set of expert judgements.

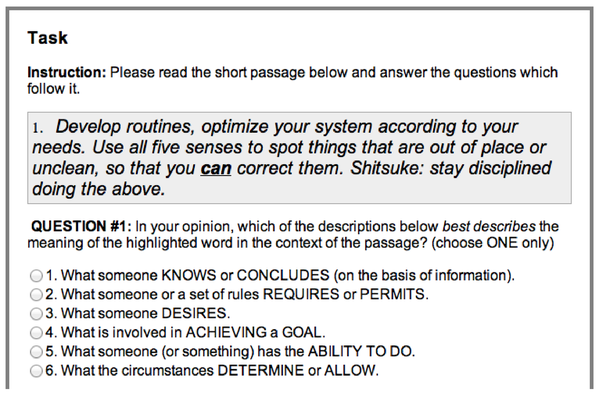
2. Methodology

2.1 Token Selection

- Tokens chosen from the UKWAC corpus: *need* (auxiliary), *can* (auxiliary), *chance* (noun), *certain* (adjective), and *likely* (adverb).
- 48 of these—around two of each type per HIT—were selected for judgement. These were randomized and fed into the Amazon HIT templates.
- Gold tokens measured baseline performance after Run 1. Adjusted corpus tokens to make easier.
- 12 of these were inserted total—two per HIT, one randomly throughout and one as the last question.

2.2 Similarities Among Runs

- Set of tokens was identical.
- Every HIT judged by five annotators.
- Identical directions, except for the third run, which featured a caveat that users who had previously completed the task would not be paid.



Task

Instruction: Please read the short passage below and answer the questions which follow it.

1. *Develop routines, optimize your system according to your needs. Use all five senses to spot things that are out of place or unclean, so that you* ***can*** *correct them. Shitsuke: stay disciplined doing the above.*

QUESTION #1: In your opinion, which of the descriptions below *best describes* the meaning of the highlighted word in the context of the passage? (choose ONE only)

- 1. What someone KNOWS or CONCLUDES (on the basis of information).
- 2. What someone or a set of rules REQUIRES or PERMITS.
- 3. What someone DESIRES.
- 4. What is involved in ACHIEVING a GOAL.
- 5. What someone (or something) has the ABILITY TO DO.
- 6. What the circumstances DETERMINE or ALLOW.

2.3 Contrast Among Runs

- Primarily by reward and length.
- Longer runs had room to bare gold standard tokens—
- The longer HITs featured choices for alternatives. (i.e., what other options did you consider?) and room for optional commentary.

Table 1: *Contrasting Runs*

	Q's per HIT	Reward	Other Feedback
Run 1	1	$0.03	No
Run 2	10	$0.20*	Alternatives and Comments
Run 3	10	$0.40*	Alternatives and Comments

*A $0.50 bonus was offered for completing all six HITs.

3. Results

89 workers participated in our study. To measure their reliability, we measured agreement of annotators with Fleiss' kappa scores. In addition to bare scores, we also looked at the improvement in scores when two sets of easily confusable categories were collapsed.

Table 2: *Gold Kappas*

Run	Region	Six Cat.	C/E Collapse	B/T Collapse	C/E and B/T	Workers (HITs per...)
2	US	0.372	0.490	0.379	0.503	17 (1.76)
3	US	0.321	0.362	0.288	0.327	28 (1.07)
3	GB	0.264	0.244	0.345	0.333	20 (1.45)

Table 3: *Corpus Kappas*

Run	Region	Six Cat.	C/E Collapse	B/T Collapse	C/E and B/T	Workers (HITs per...)
1	US	0.099	0.154	0.122	0.183	24 (5.55)
2	US	0.180	0.249	0.201	0.277	17 (1.76)
3	US	0.206	0.206	0.220	0.222	28 (1.07)
3	GB	0.225	0.323	0.247	0.352	20 (1.45)

Selected Sets of Kappa Scores

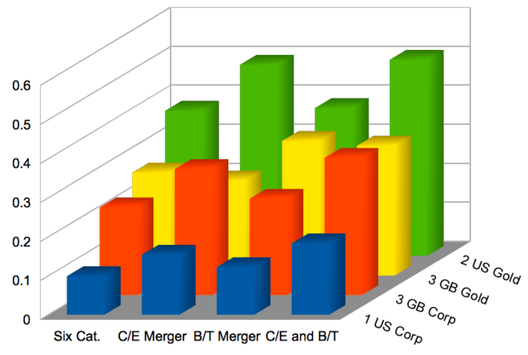


Table 5: *Confusion Matrix of Runs 2 and 3*

		Turker Judgements											
		ability		bouletic		circumstantial		deontic		epistemic		teleological	
Expert	ability	50	(56.2%)	0	(0.0%)	7	(7.9%)	1	(1.1%)	17	(19.1%)	14	(15.7%)
	bouletic	7	(25.0%)	13	(46.4%)	1	(3.6%)	0	(0.0%)	6	(21.4%)	1	(3.6%)
	circumstantial	70	(23.7%)	11	(3.7%)	85	(28.8%)	19	(6.4%)	88	(29.8%)	22	(7.5%)
	deontic	4	(13.3%)	2	(6.7%)	4	(13.3%)	16	(53.3%)	3	(10.0%)	1	(3.3%)
	epistemic	11	(6.2%)	7	(3.9%)	34	(19.1%)	12	(6.7%)	111	(62.4%)	3	(1.7%)
	teleological	1	(0.7%)	22	(14.7%)	15	(10.0%)	35	(23.3%)	13	(8.7%)	64	(42.7%)
	Totals (Turker)	143		55		146		83		238		105	

4. Fleiss' Kappa

- A measure of how much better agreement is than random agreement.
- A generalization of Cohen's kappa, which only worked for two annotators.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (1)$$

- $\bar{P}$ is the average of how much a set of raters agree on a subject,
- $\bar{P}_e$ is the sum of squares of the proportions of choices made within each modal type.
- The numerator represents degree of agreement over random
- The denominator represents the highest degree of agreement possible better than chance. [4]

The very definition of Fleiss' kappa gives meaning to 0 and 1 on the scale, but where exactly good agreement begins is a difficult question [3] [4]. Ng (1999) used kappa scores to measure inter-annotator agreement on a word sense disambiguation task [5]. Other scales have been developed to make meaningful interpretations of kappas. [2]

Table 6: *Examples of Kappa Scores*

Significance	Source	Score
Random	[3]	0
Worst Raw Corpus	Run 1	0.099
Avg. Raw	[5]	0.317
Best Collapsed Gold	Run 2 US	0.503
Tentative Result	[2]	0.670
Good Reliability	[2]	0.800
Best Collapsed	[5]	0.862
Perfect Agreement	[3]	1

5. Discussion

Overall it seems:

- Our kappa scores are low. Although agreement above random is interesting, a truly strong result is significantly higher than just random . However, many of the corpus tokens are categorically ambiguous. Kappa scores, although measuring inter-annotator agreement, do not measure patterned disagreement. Also, within the subset of corpus kappas, there was dramatic improvement after lessons from Run 1 were applied to future runs.
- Turkers seem to match the expert category about half the time, excluding circumstantial. Circumstantial is confused with epistemic; collapsing these dramatically increases kappa scores.

This work was funded by NSF grant number BCS-1053038.

References

[1] Artstein, R., Poesio, M. Inter-Coder Agreement for Computational Linguistics. Extendended version of article published in *Computational Linguistics*. 2007.

[2] Carletta, J. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*. Vol. 22, No. 2, June 1996.

[3] Cohen, J. A Coefficient of Agreement for Nominal Scales *Educational and Psychological Measurement* Vol. 20, No. 37, 1960.

[4] Fleiss, J. L. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*. Vol. 76, No. 5, 1976.

[5] Ng, H.T., Lim, C.Y., Foo, S. K. A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation. *Proceedings of the Siglex-ACL Workshop on Standarizing Lexical Resources*. 1999.