Narrative Cloze Task: Experiment 1

Erin Bennett
erindb@stanford.edu

Experiment

The narrative cloze task is a common evaluation task for script induction. I would like to understand humans' performance on the narrative cloze task.

In this task, participants read stories involving a particular entity. One of the events the entity is involved in is left blank. Participants are asked to fill in this blank with a plausible completion.

Demographics and attention checks

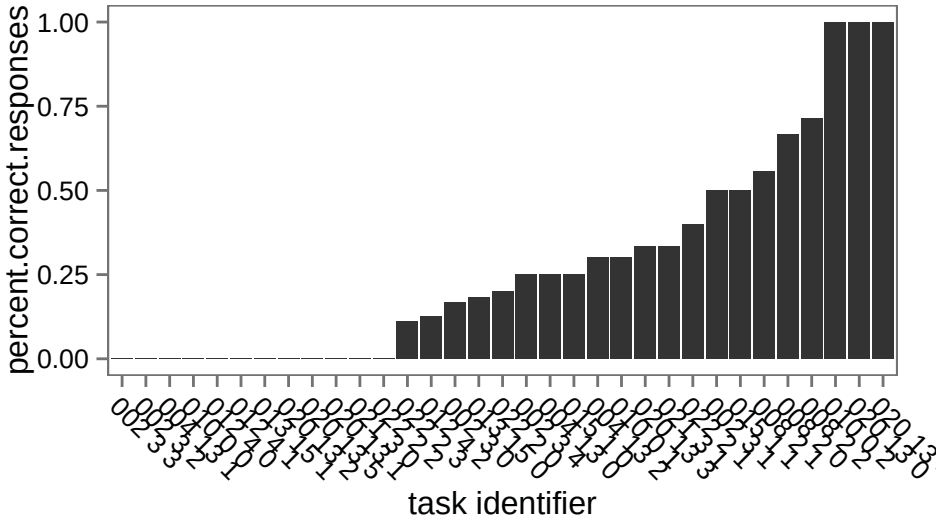
Average completion time was 6 minutes.

3 participants were excluded for failing an attention check.

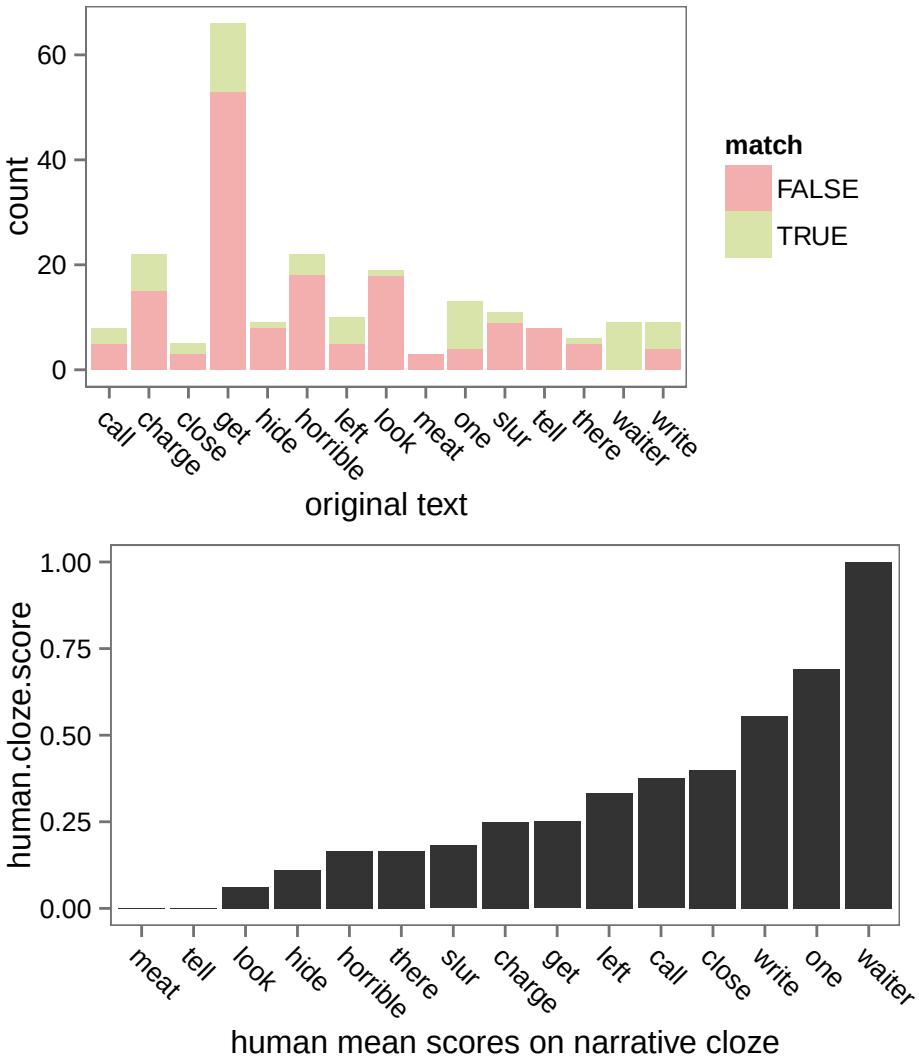
Exploratory results

How do people do on the narrative cloze task?

The average percent correct responses on a narrative cloze task is 28% (when using synonyms and paring down to minimal caveman speak).



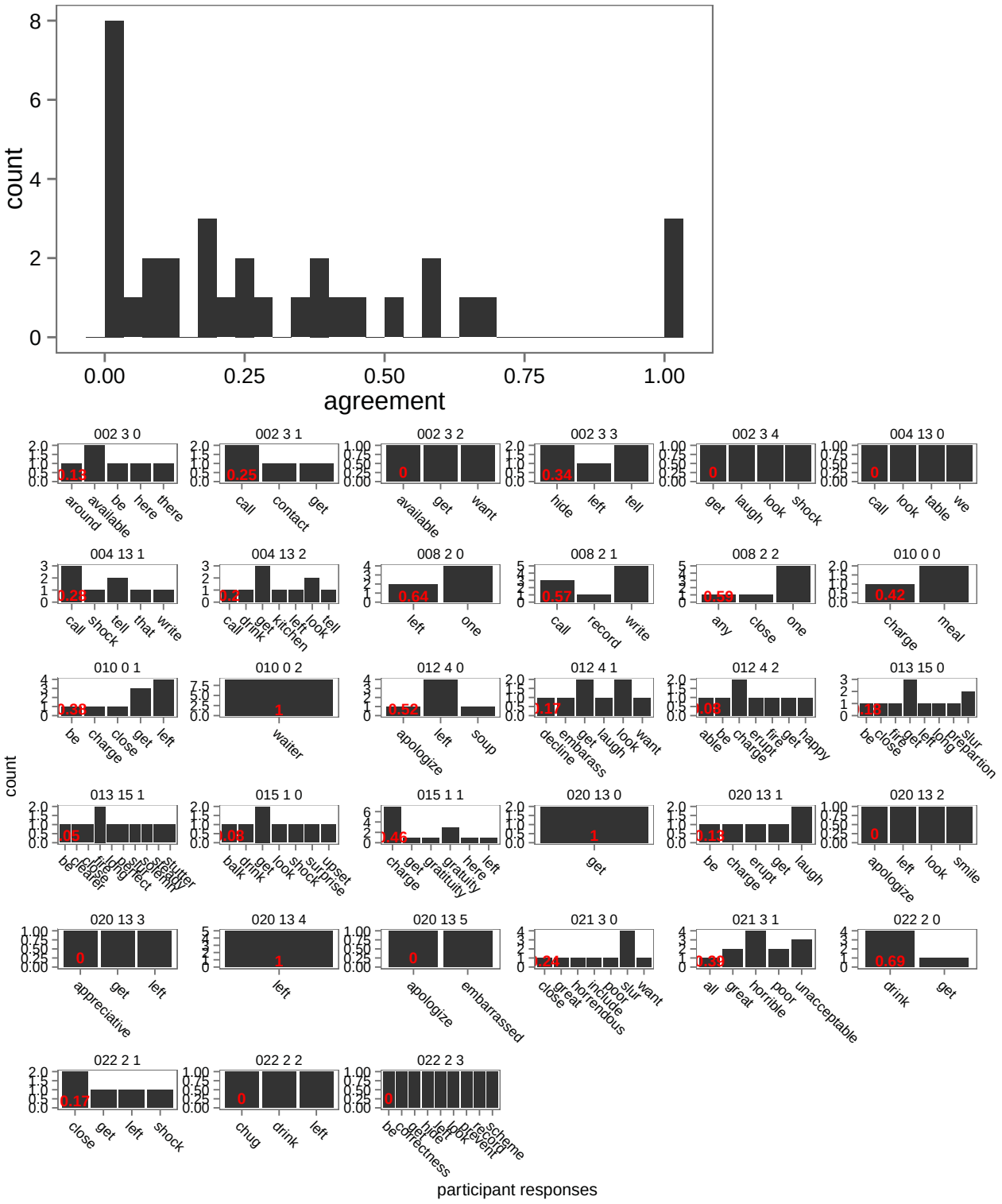
Some words are easier for people to fill in than others (several cloze tasks had the head word 'get', which is why there are so many).



How much do people tend to agree with one another?

I'm taking the "entropy" of the distribution that people give back and dividing by `n` as a measure of disagreement. So a completely uniform distribution where every participant gives a different response from every other participant will have disagreement value of 1 and an agreement value of 0. A distribution where every participant gives the same response will have an agreement value of 1.

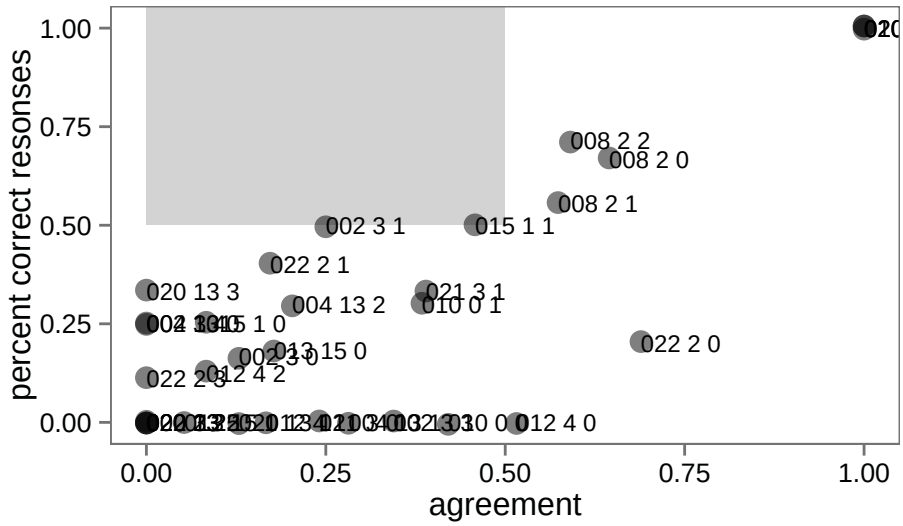
Across all the cloze tasks, agreement is peaked at 0.



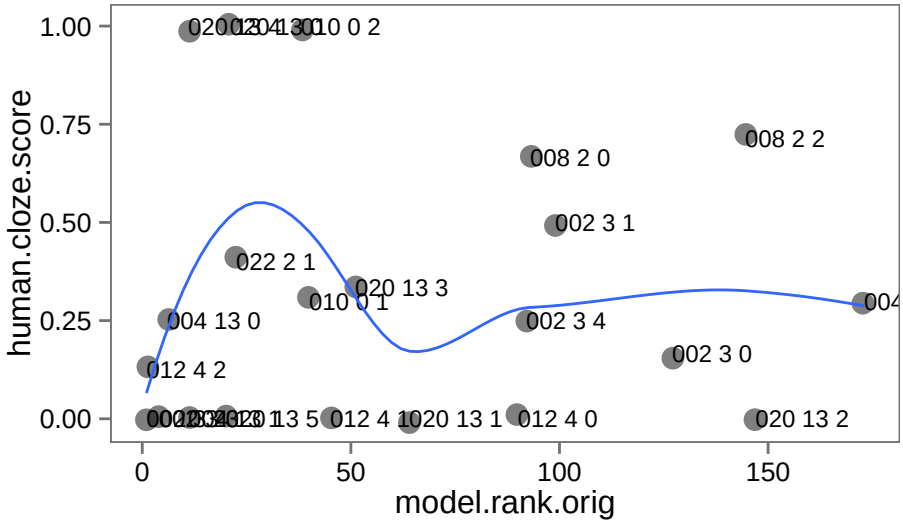
How does agreement among people relate to narrative cloze performance?

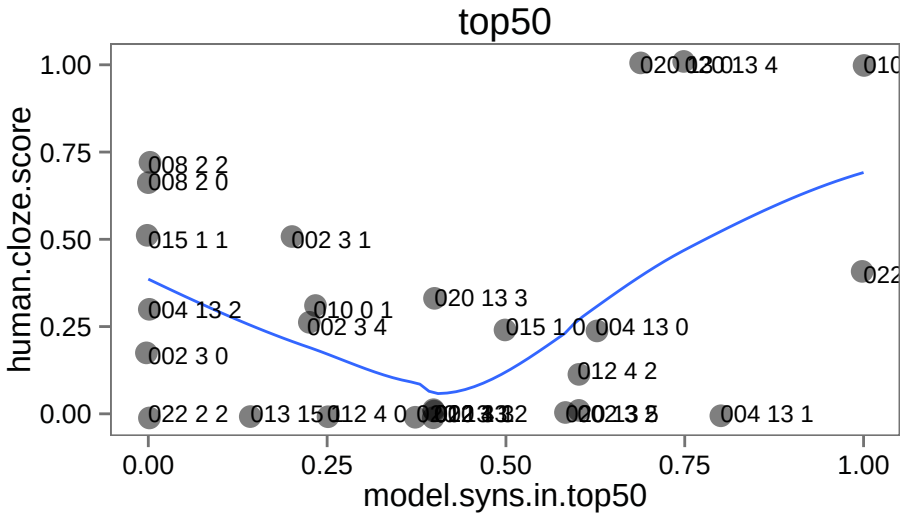
Obviously if people don't agree with one another, then the average score for that cloze task will not be very high.

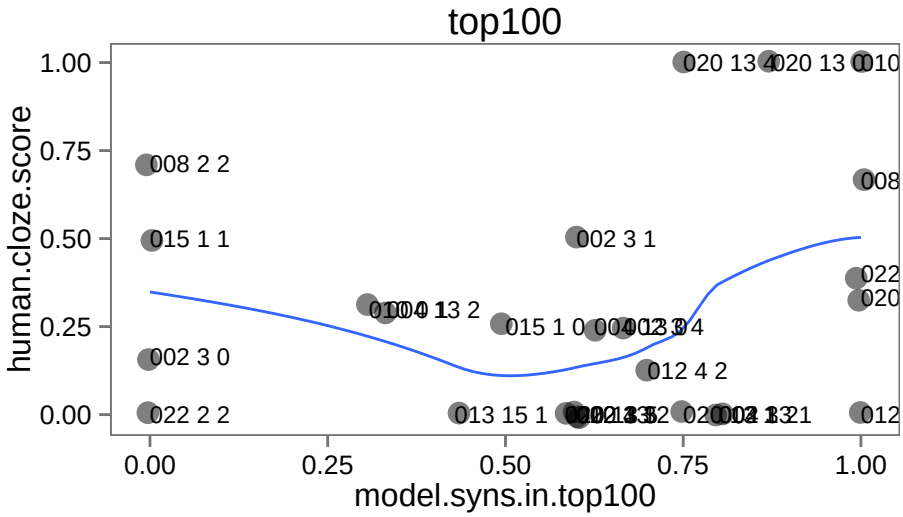
I grayed out the upper-left region of this graph to show it's basically impossible.

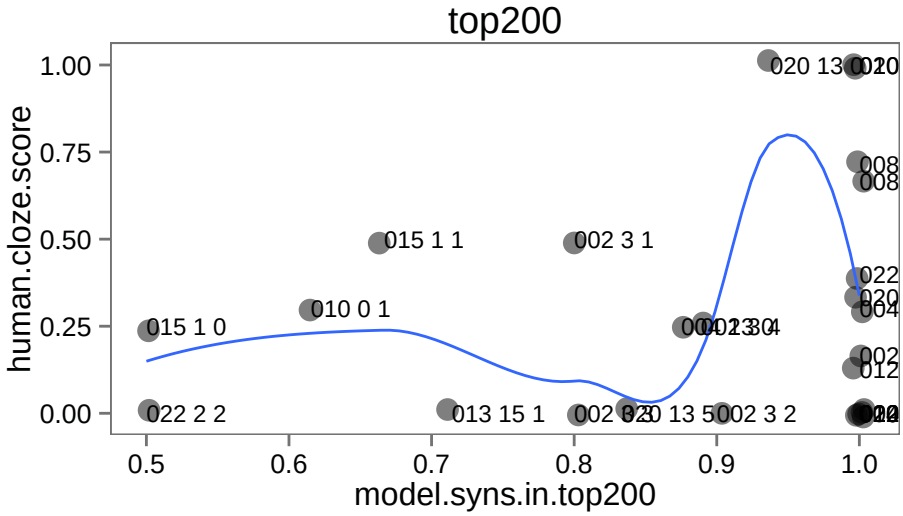


How does the model perform relative to people?






top100
human.cloze.score
model.syns.in.top100
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
020 13 4
020 13 010
008 2 2
008
015 1 1
002 3 1
022
020
002 3 0
012 4 2
022 2 2
013 15 1
012


top200
human.cloze.score
model.syns.in.top200
1.00
0.75
0.50
0.25
0.00
0.5
0.6
0.7
0.8
0.9
1.0
020 13
008
008
015 1 1
002 3 1
022
020
010 0 1
015 1 0
022 2 2
013 15 1
002 3 2

References