Narrative Cloze Task: Experiment 3

Erin Bennett
erindb@stanford.edu

Background

Scripts are a kind of knowledge that people have that helps them understand narrative texts, including filling in their representation of the story with events that were not mentioned but that probably happened.

Recent NLP work has developed initial script induction models and evaluated their performance at filling in the gaps in narrative text. Some of these models have access only to very pared-down linguistic information (essentially only the main verb of a sentence), which allows them to combine multiple sentences together but potentially looses a lot of important information, too.

Question

How do people fill in gaps in narrative text?

- To what extent can people accurately fill in gaps in natural narrative text?
- Whether or not people are able to fill in the original text, do they tend to agree with one another?
- How detailed does the contextual linguistic information need to be for people to sensibly fill in the gaps in narrative texts?

And how does this relate to the performance of script induction models?

Experiment

Design

51 Stimuli:

- 17 documents randomly sampled from "Dinners from Hell" blog (within Ss)
- 3 event chain / cloze test pairs for each document (between Ss)

3 linguistic conditions (between Ss):

- caveman
- event only
- original text

Pilot results

In the pilot, we lost data from some participants, and unfortunately, this was length/time based and so the most conscientious participants were excluded.

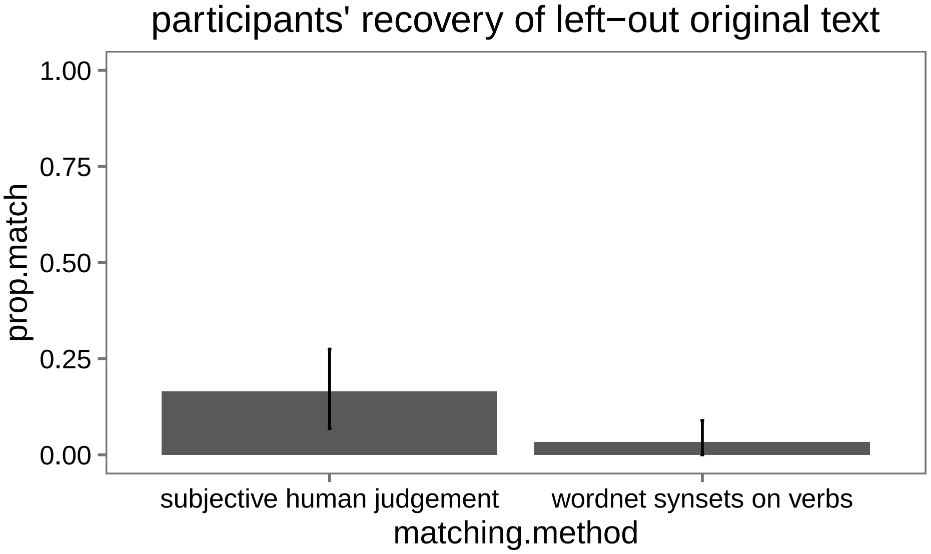
There were 51 cloze tasks in total, and only 17 were seen by each participant. We collected data from 24 participants, so not all of the cloze tasks had enough responses to make meaningful inferences about (e.g.) response overlap across participants.

Question 1: To what extent can people accurately fill in gaps in natural narrative text?

First, we wanted to know whether the main verb of the original sentence was mentioned in the participants' responses. We searched wordnet for synonyms of each of the verbs that participants used checked for any overlap with the synonyms of the original main verb.

Of the 41 stimuli presented with the original passage text as context, only 2 of them had main verbs that were shown to be recovered (i.e. shared a synonym with one a response verb) by at least one participant.

However, hand-coding whether the gist of the response was similar to the gist of the original event, we can see that people are actually doing a lot better on this task. Still, only about 0.25% of the stimuli were recovered by any of the participants, and on those conditions, only 0.658% of participants chose that response.



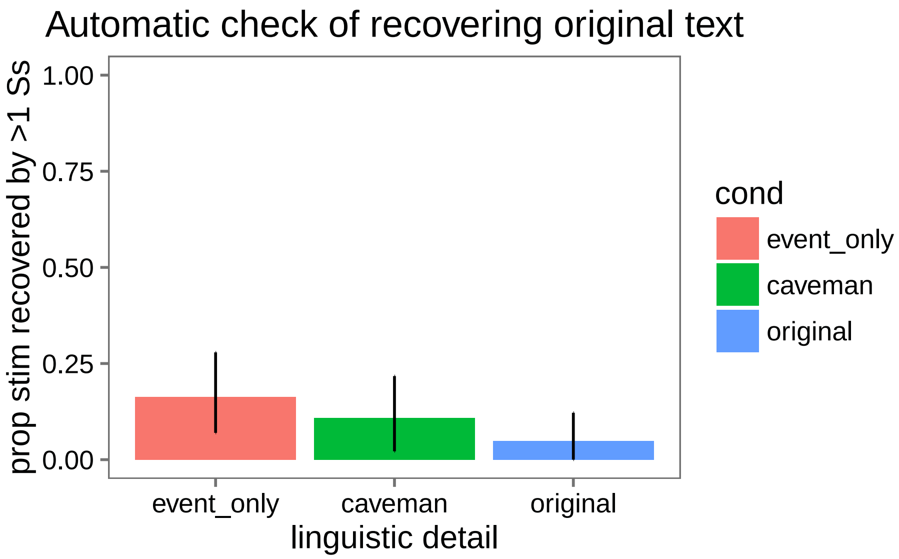
Overall, people don't seem to perform especially well on the task of recovering left-out sentences.

Question 2: Whether or not people are able to fill in the original text, do they tend to agree with one another?

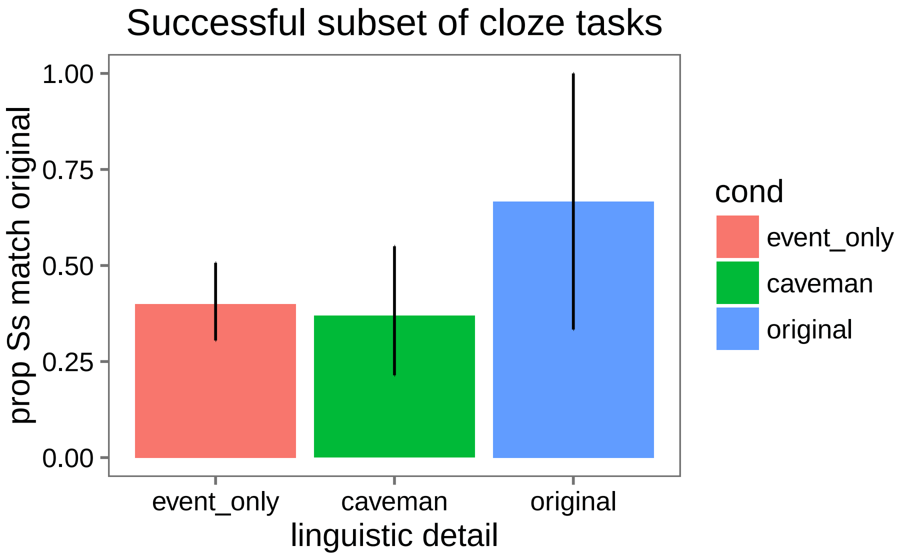
Question 3: How detailed does the contextual linguistic information need to be for people to sensibly fill in the gaps in narrative texts?

The other two linguistic conditions did not differ much in the proportion of participants who recovered the original text.

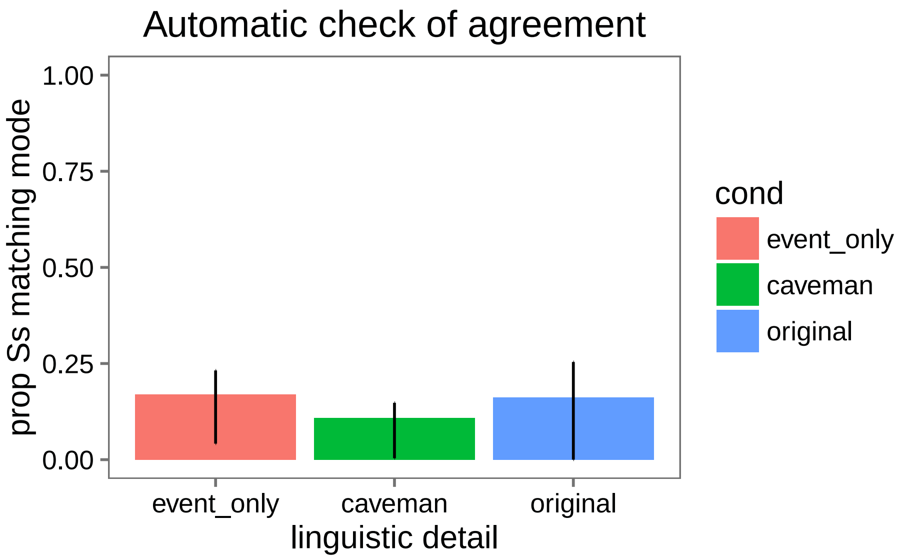
Numerically, the event-only condition resulted in higher rates of successful recovery of the original text.

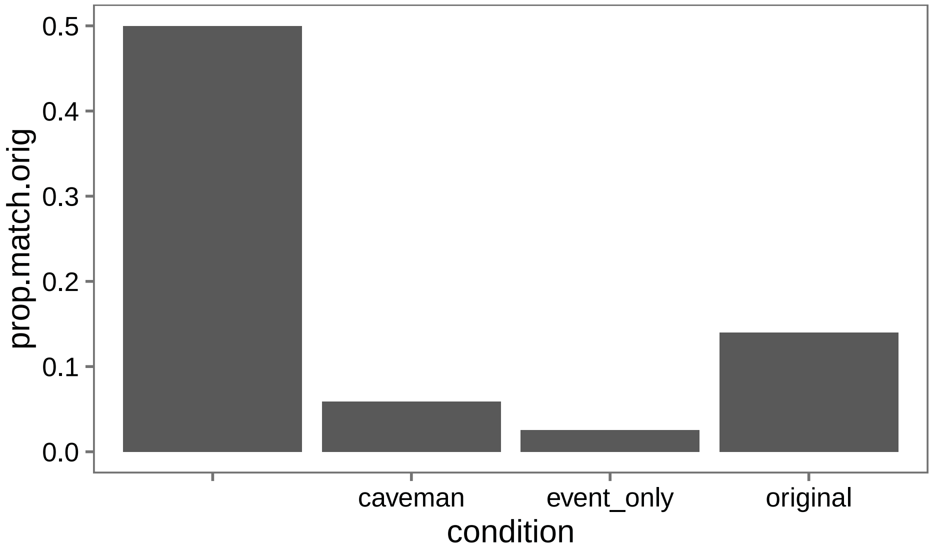


Among *only* the stimulus/condition pairs where at least one person recovered the original, a numerically higher proportion of participants recovered the original text when the context was provided in full linguistic detail. But there were only two tasks in that condition where any participants succeeded.



In each of the three linguistic detail conditions, the most common response verb was given by around 20% of participants. The three conditions did not differ much.





histograms and entropy link in weekly update with additional figures is there anything in the data that suggests people are being informative?

Question 4: How does human performance on this task relate to the performance of a PMI-based script induction model?

Future work

One important observation is that most things that are mentioned in a narrative are at least a little bit uncommon given the context, otherwise they would not be interesting/informative enough to mention. How does this affect what people and machines can learn from narrative text?

References