INTRODUCTION

- Discerning plausible from implausible events: crucial building block for NLP
- Previous work mostly focused on semantic knowledge to distinguish
  - physically plausible vs. implausible events
  - events with mostly conceptually concrete participants

RESEARCH GOALS & CONTRIBUTIONS

- Create novel dataset for physical and abstract plausibility of events in English,
capturing abstractness to the same extent as concreteness for the first time
- Systematically examine plausibility across levels of abstractness
- Explore and represent disagreement in plausibility annotation

CONSTRUCTING EVENT TARGETS

PLAUSIBLE EVENTS (marked in blue)
- From English Wikipedia sample: Extract attested events, filter for profanity, assign abstractness ratings, bin according to abstractness, and sample 1,080 plausible events

(PSEUDO)-IMPLAUSIBLE EVENTS (marked in yellow)
- Based on extracted attested triples:
  - (i) Automatically generate pseudo-implausible events by perturbating event constituents
  - (ii) Construct 1,080 pseudo-implausible event similarly to plausible event construction

CAPTURING (SEMANTIC) PLAUSIBILITY

PLAUSIBILITY
- Captures non-surprise in a given context child-sleep vs. tree-sleep
- Includes both what is preferred (and probably most plausible) and what is unusual (but still very much plausible), child-eat-banana vs. child-eat-pebble
  → in contrast to selectional preference / thematic fit
- Can be estimated as a matter of degree with events assessed corresponding to perceived plausibility child-eat-banana vs. child-eat-pebble vs. child-eat-skyscraper
- Denotes what is likely in a given world but not necessarily attested in a given corpus human-dies vs. human-breathes

COLLECTING HUMAN ANNOTATIONS

TASK: Collect plausibility judgements on AMT for 2,160 plausible and implausible triples

Approval Rate ≥ 98% + ≥ 1k appr. HITs
Discard responses from workers with...
Quality checks

DATASET STATISTICS
- 15,571 plausibility ratings for 1,733 triples
- Ø IAA: Soft Jaccard Coefficient of 0.64
  → reasonable agreement among annotators with indication of disagreement to be examined

ANALYSIS OF HUMAN JUDGEMENTS AND DISAGREEMENT

What can we learn from rating distributions?

- Humans tend to favor plausibility over implausibility, while avoiding the extreme on the plausibility end of the scale.
- Implausibility yields higher disagreement as annotators disagree more when rating triples originally labelled implausible.

CONCLUSIONS

- Presented a novel human-annotated dataset for physical and abstract plausibility for events in English
- Explored relationship between abstractness and plausibility and analyzed annotator disagreement
- Released both raw and a range of aggregated annotations to foster research on (semantic) plausibility and related notions, disagreement, and relevant downstream tasks such as commonsense reasoning

ACKNOWLEDGEMENTS

- Presented a novel human-annotated dataset for physical and abstract plausibility for events in English
- Explored relationship between abstractness and plausibility and analyzed annotator disagreement
- Released both raw and a range of aggregated annotations to foster research on (semantic) plausibility and related notions, disagreement, and relevant downstream tasks such as commonsense reasoning