# Evaluating state-of-the-art models of speaker commitment

Nanjiang Jiang and Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University

THE OHIO STATE UNIVERSITY

# Speaker Commitment

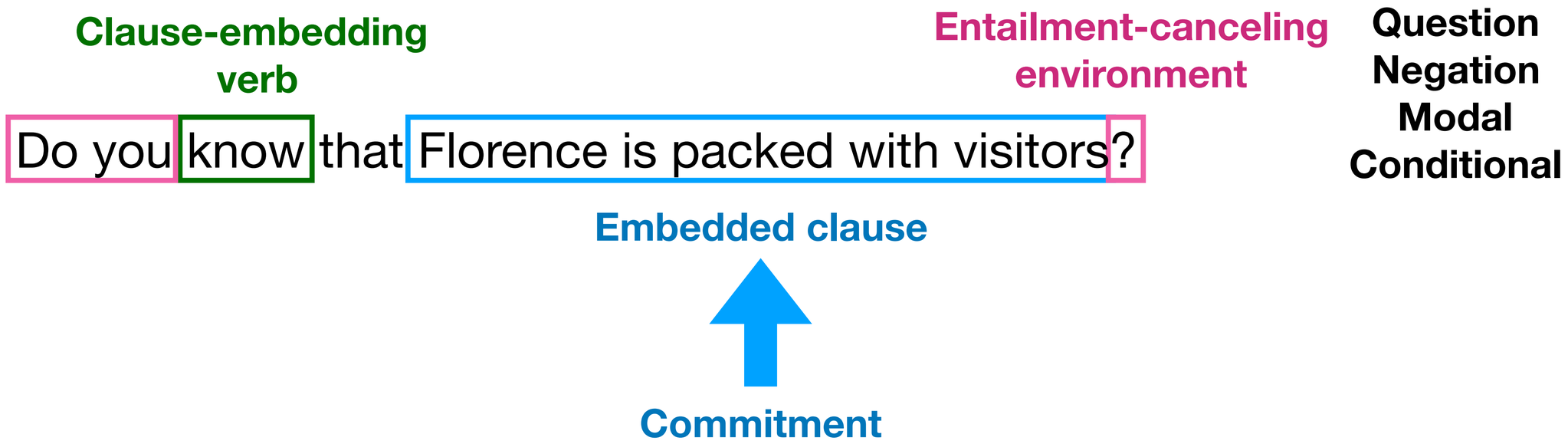
How certain a speaker is that the event in their utterance is a fact?

Florence is the most beautiful city I’ve ever seen.

Florence might be the most beautiful city I’ve ever seen.

# Speaker Commitment

How certain a speaker is that the event in their utterance is factual?



# Speaker Commitment

How certain a speaker is that the event in their utterance is factual?

**Factive verb** Karttunen 1971

Do you know that Florence is packed with visitors?

Florence is packed with visitors.

# Speaker Commitment

How certain a speaker is that the event in their utterance is factual?

**Factive verb**

Do you know that Florence is packed with visitors?

**Nonfactive verb**

Do you think that Florence is packed with visitors?

Florence is packed with visitors.

# Speaker Commitment

How certain a speaker is that the event in their utterance is factual?

**Factive verb**

Do you know that Florence is packed with visitors?

**Nonfactive verb**

Do you think that Florence is packed with visitors?

**Neg-raising**

I don’t know that Florence is packed with visitors.

Florence is packed with visitors.

# CommitmentBank

**Mary:** ***I said you were mad to come over at this time. It's a world event. Do you know that Florence is packed with visitors?***

How certain is Mary that Florence is packed with visitors?

| ◎ | ◎ | ◎ | ◎ | ◎ | ◎ | ◎ |
|---|---|---|---|---|---|---|
| 3 | 2 | 1 | 0 | -1 | -2 | -3 |
| Mary is certain that it is true | | | Mary is not certain whether it is true or false | | | Mary is certain that it is false |

**8+ annotations/item**

**556 items for our evaluation**

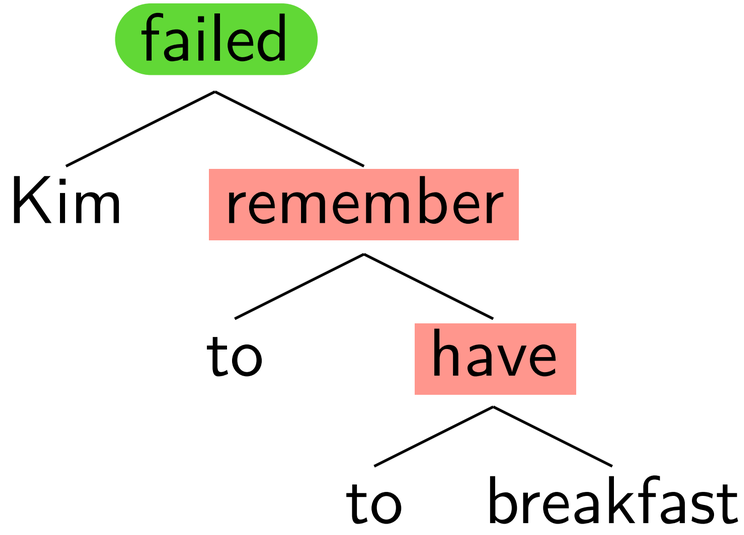
**de Marneffe et al 2019**

# Models

1. Rule-based TruthTeller by Stanovsky et al 2017
2. BiLSTM-based models by Rudinger et al 2018

# Rule-based TruthTeller

Kim failed to remember to have breakfast.



Kim had breakfast.

Stanovsky et al 2017, Lotan et al 2013

Nairn et al 2006, Karttunen 2012

# Bi-LSTM models

3 typologies

- linear
- tree
- hybrid

4 datasets multitask training

- FactBank
- UW
- Meantime
- IH

Rudinger et al 2018

# Metrics

Predicting gradience of commitment $\in [-3, 3]$

**Pearson R Correlation**

**[-1, 1]**

Variability

**Mean Absolute Error**

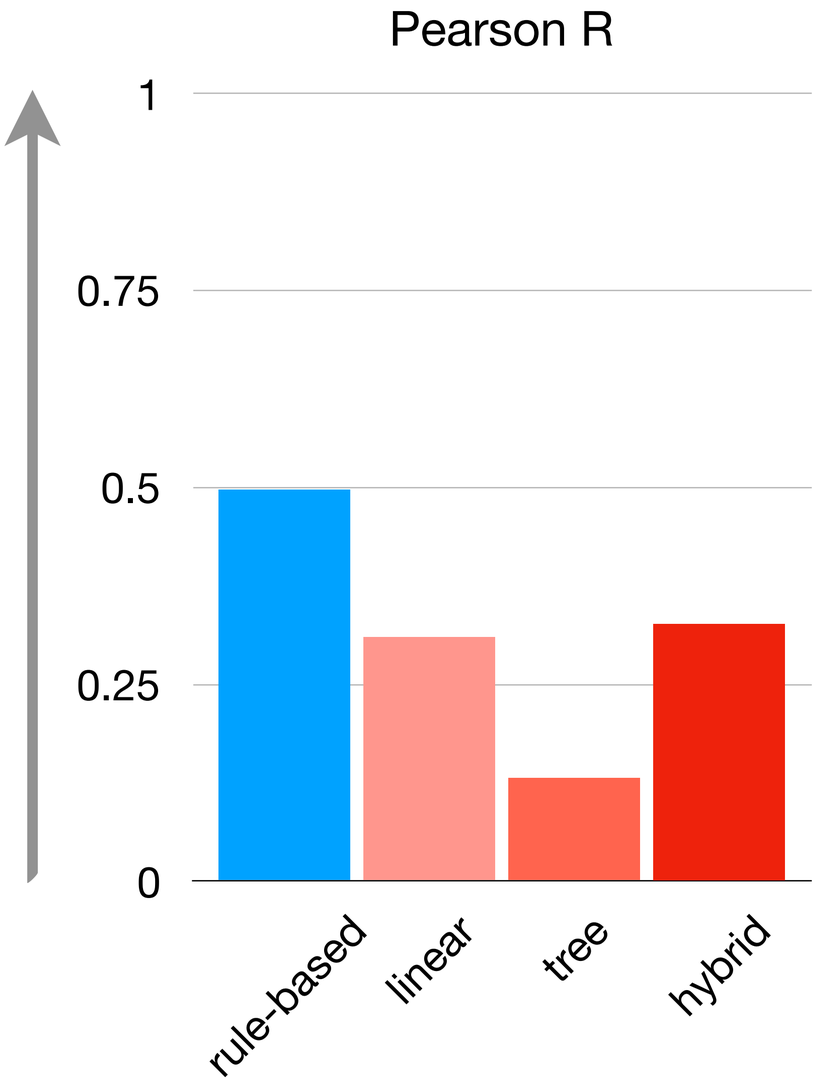
**[0, 3]**

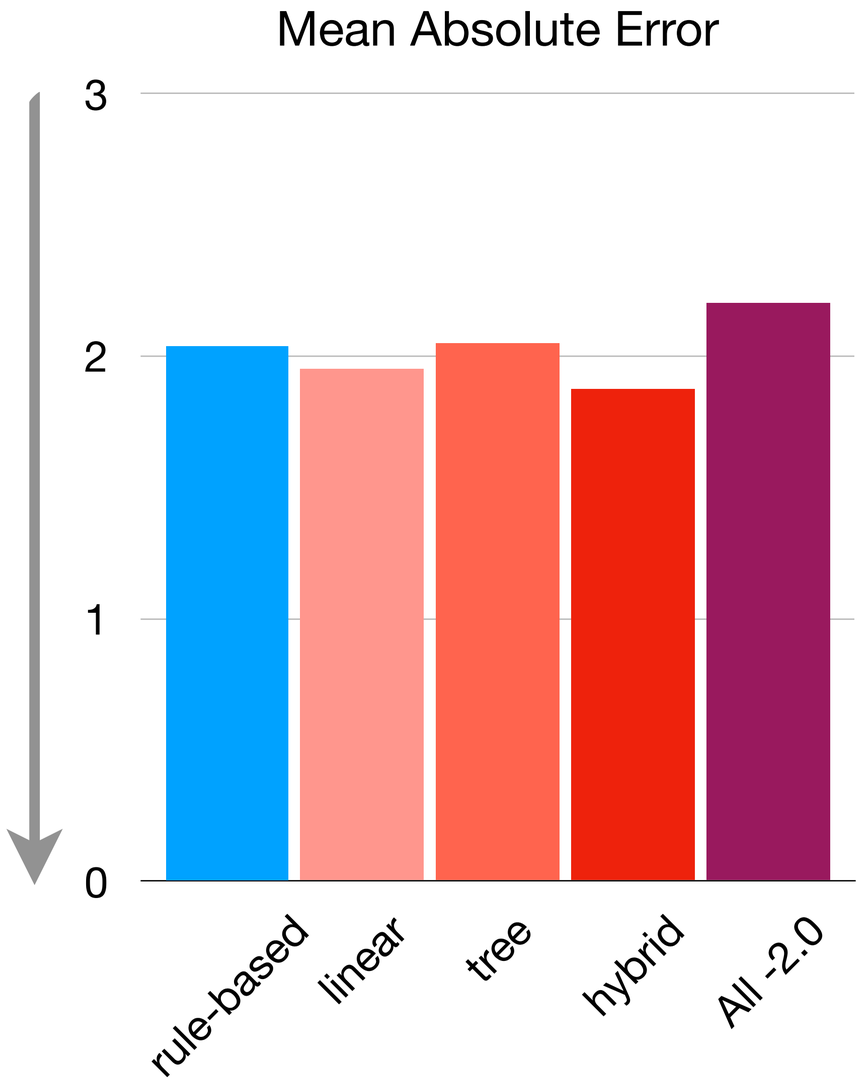
Absolute fit

**More informative for biased datasets**

# Results

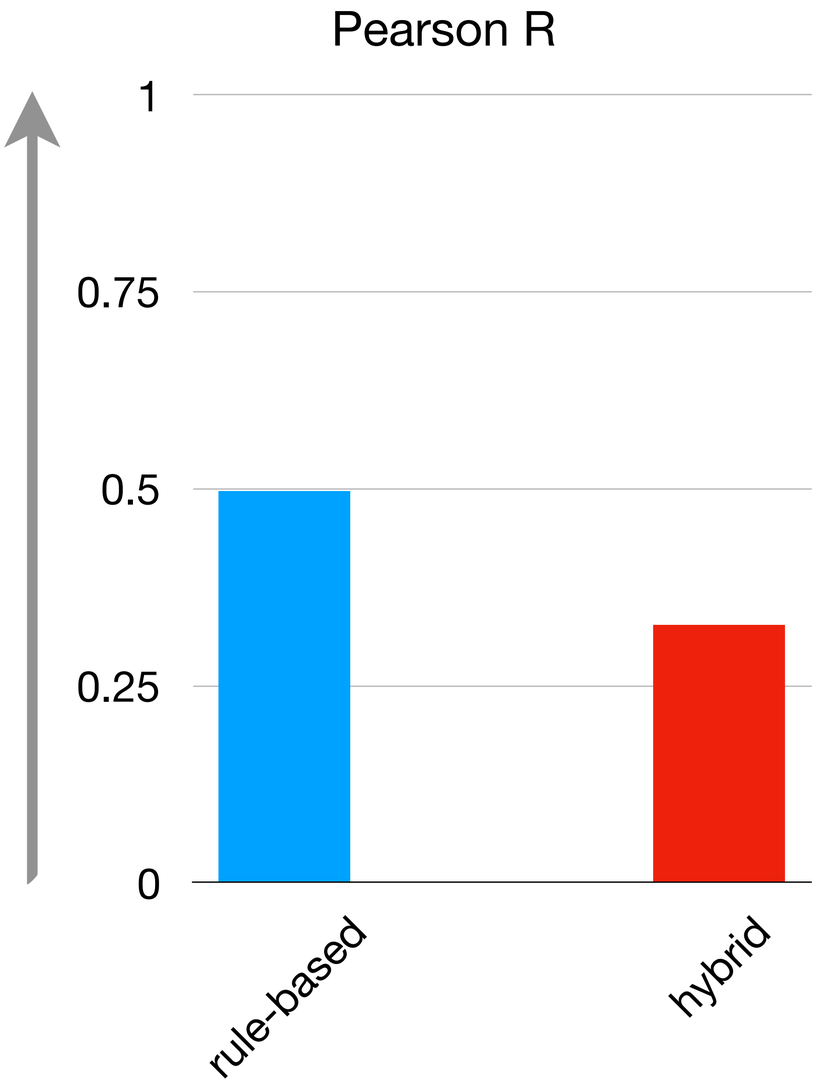
## Performance is not great overall

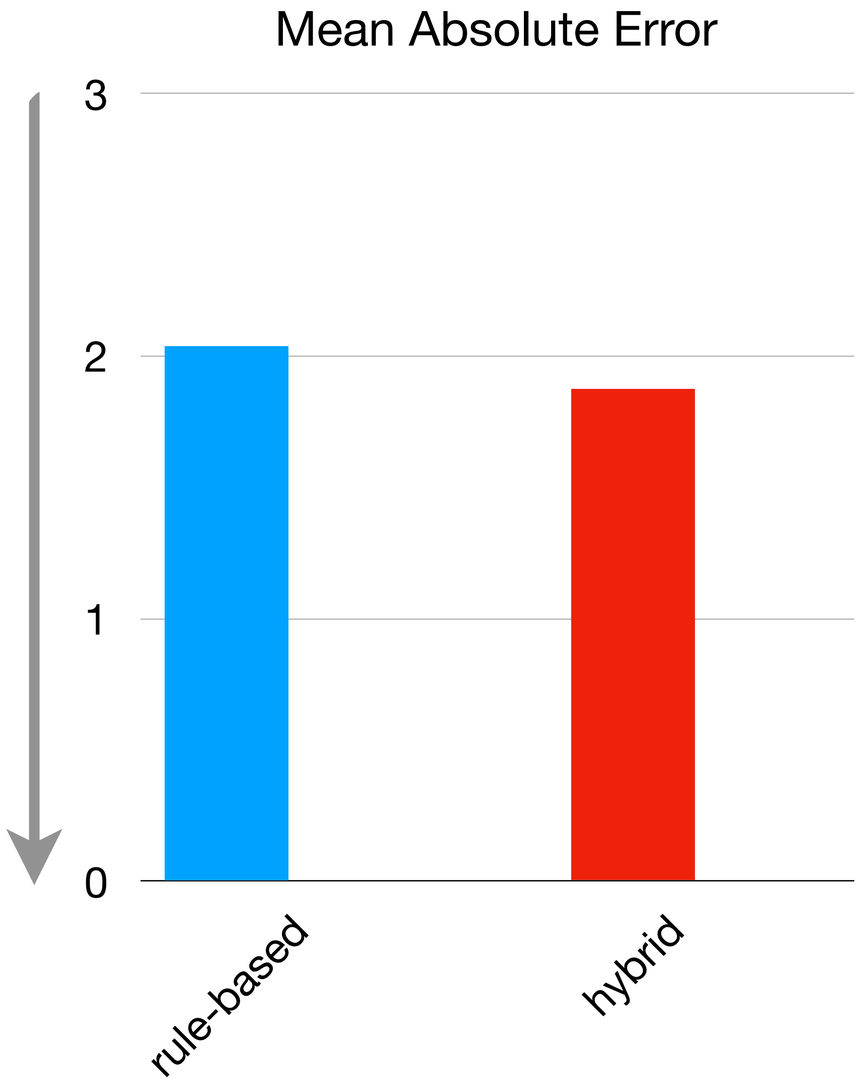


# Analysis

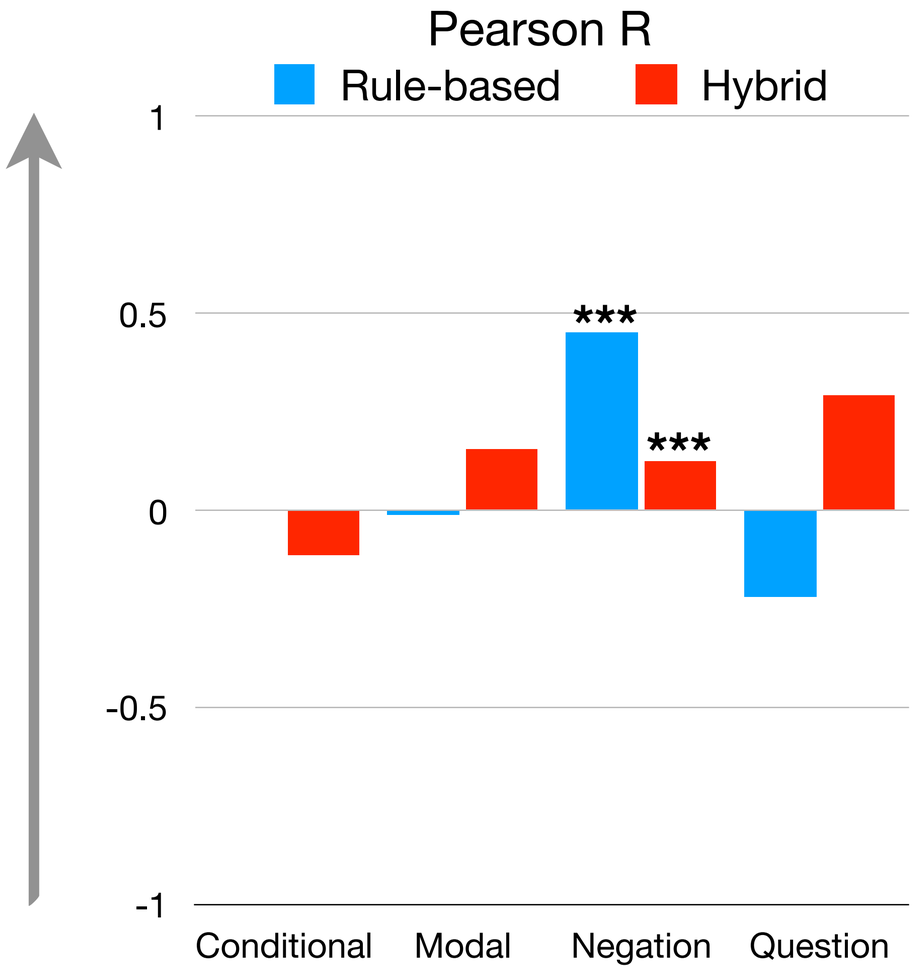
## Focus on the best performing models

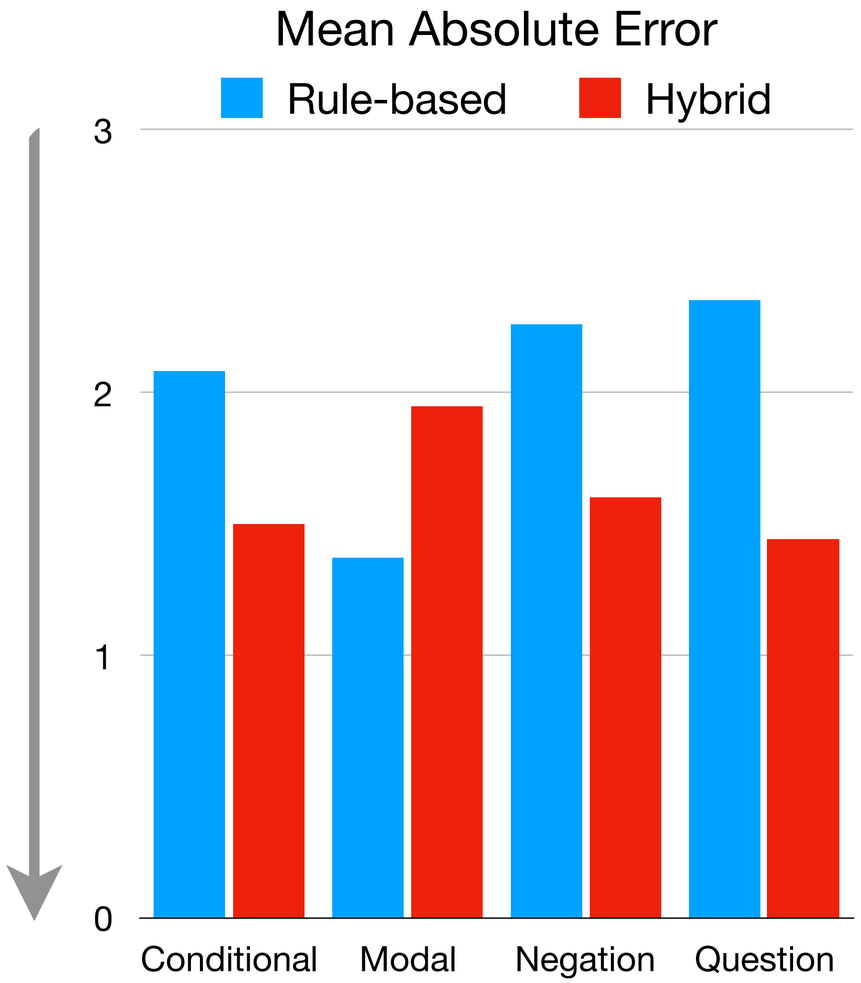


# Entailment-canceling Environment

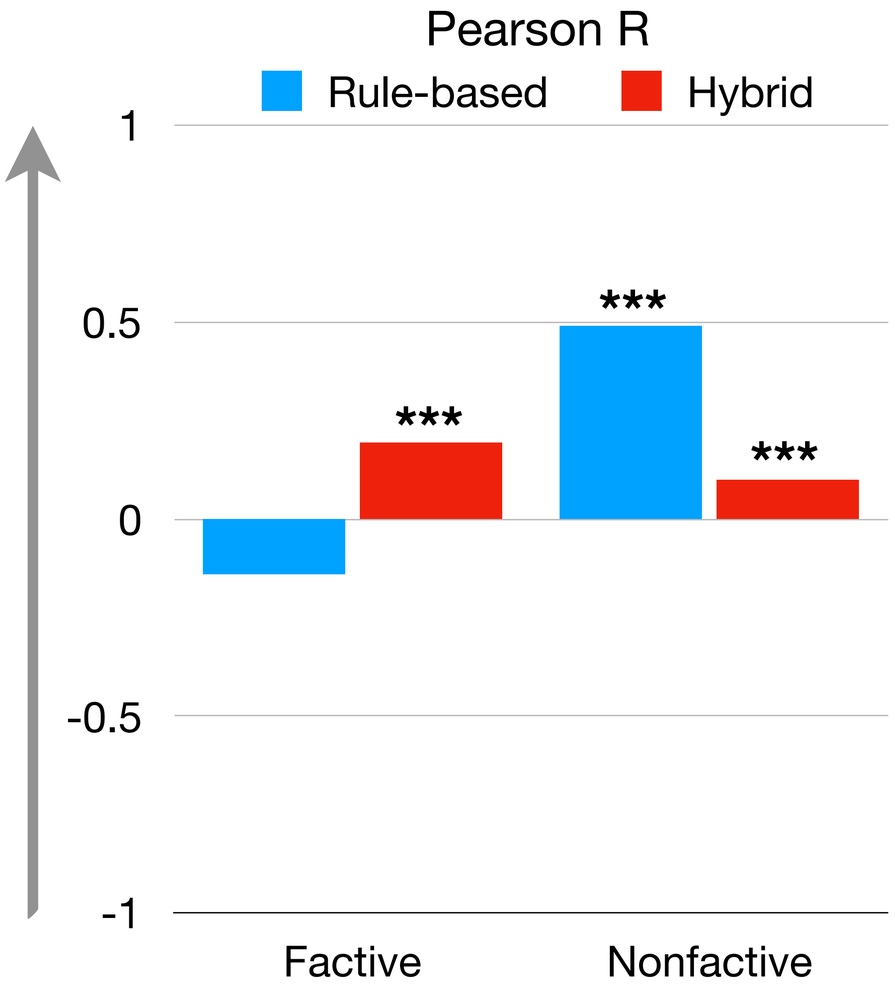
## Negation the best, conditional the worst

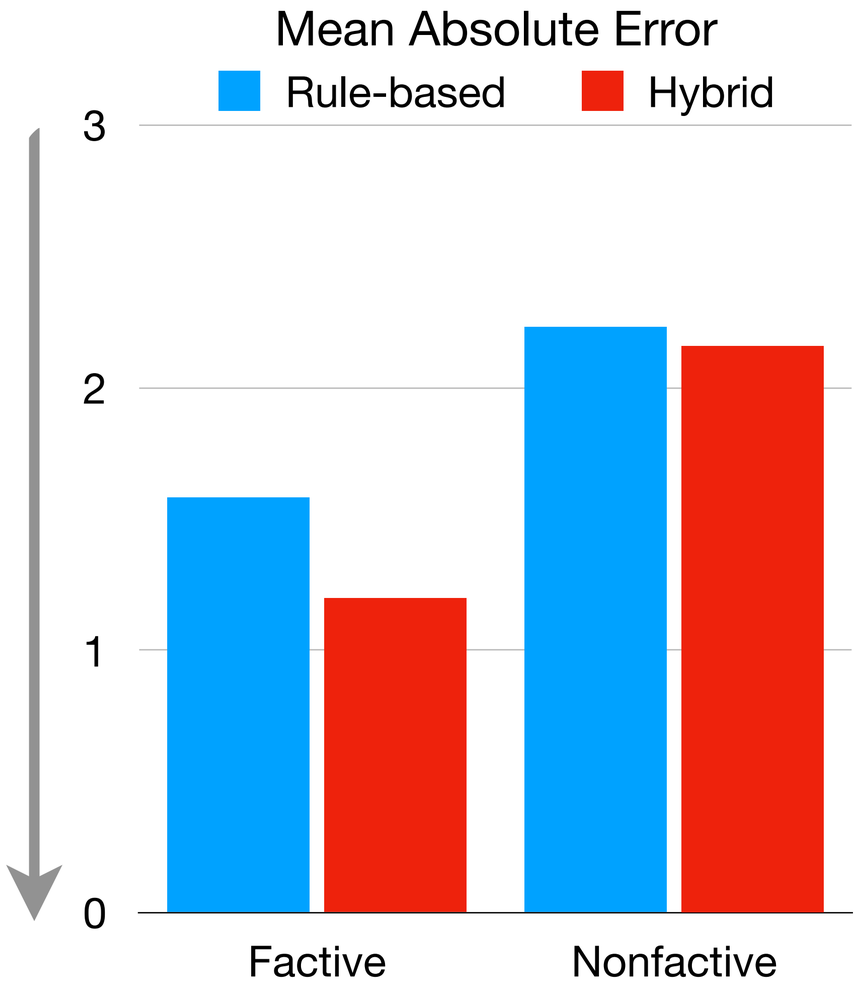


# Factivity

## Good absolute fit but poor correlation for factive verbs



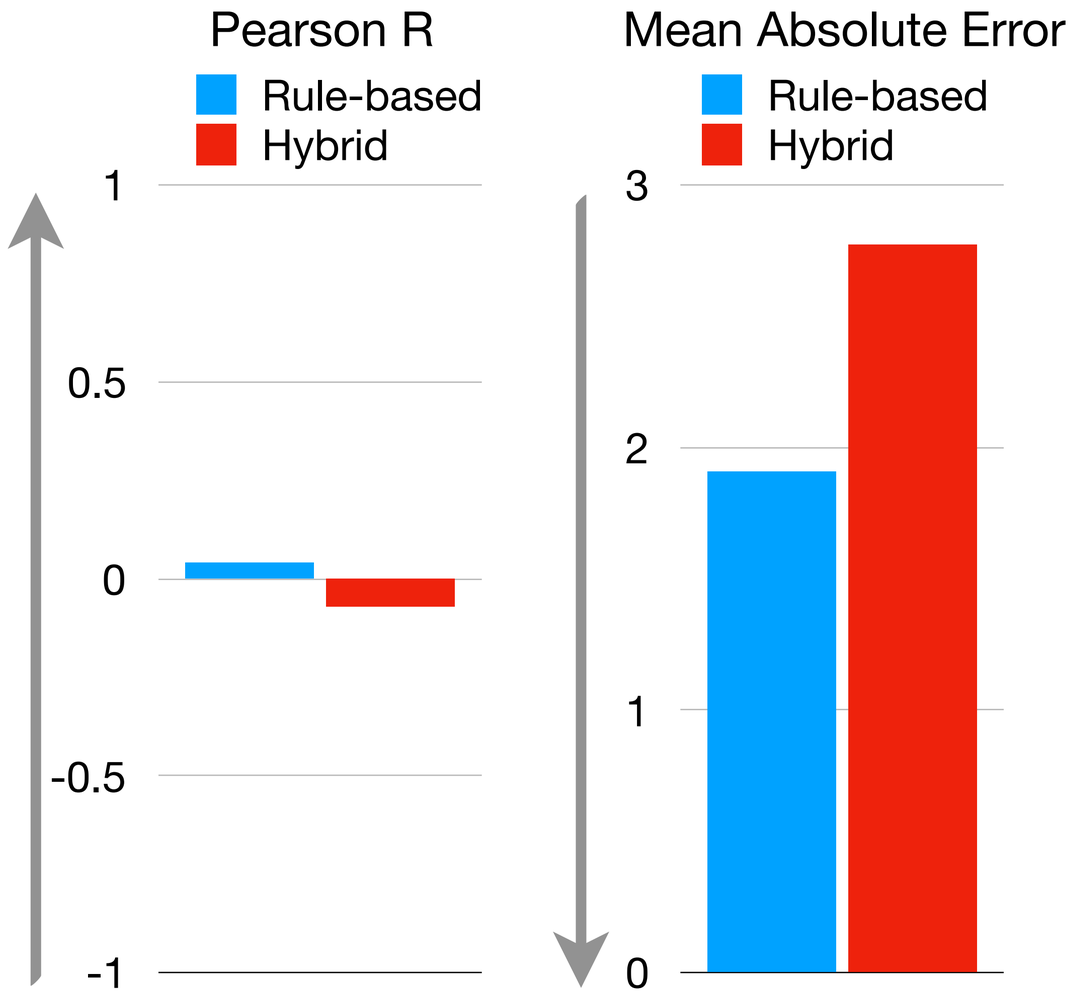

Beaver 2010, de Marneffe et al 2019

# Neg-raising

## Neither model can handle

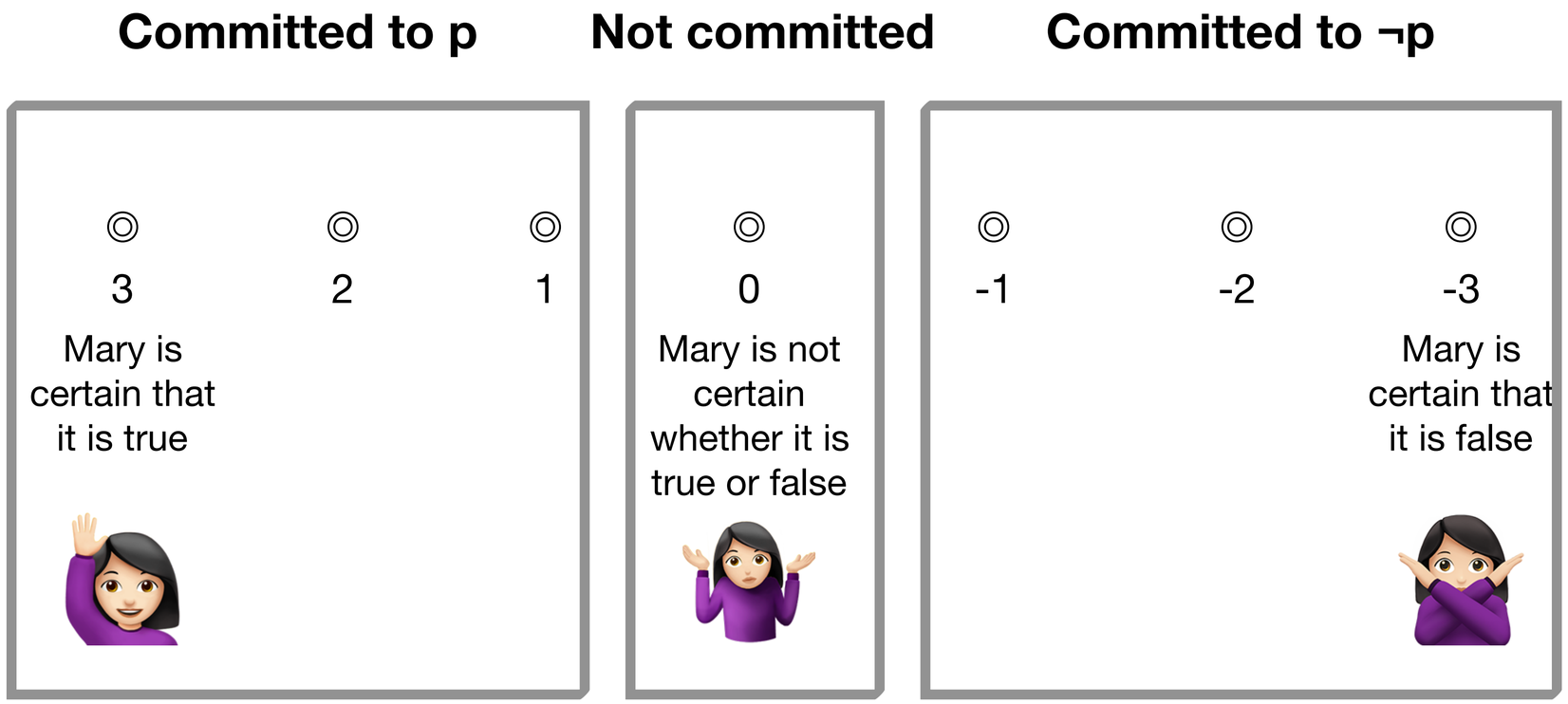
I don’t think that’s a good idea.

🙅🏻‍♀️ That’s a good idea.



# Commitment Polarity

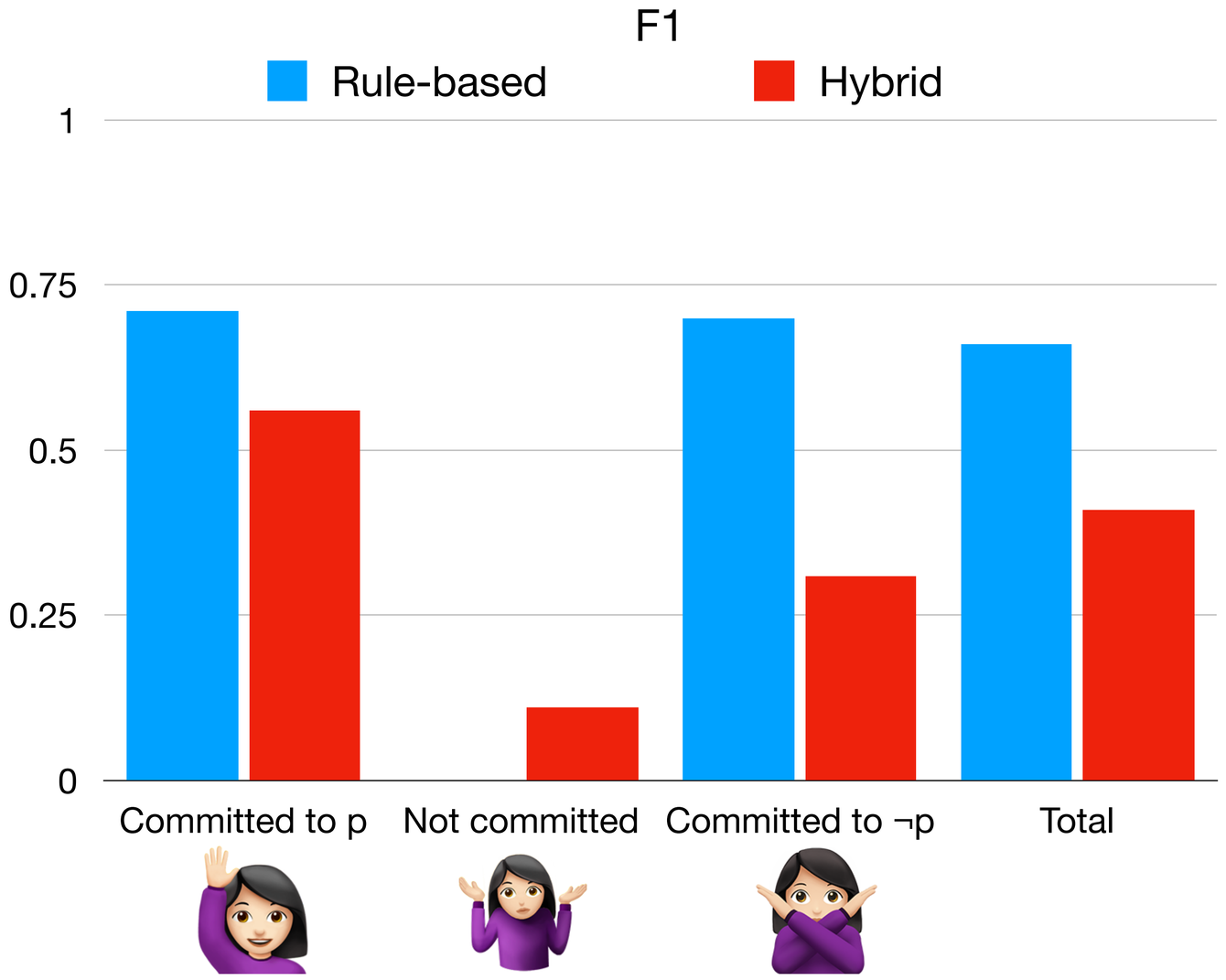
## 3-way classification




de Marneffe et al 2019

# Commitment Polarity

## Neither model can identify no commitment



# Conclusion

Conditionals, factives verbs, neg-raising are hard for models

Models can identify polarity but not gradience

Linguistically motivated models scale more successfully

# Thanks to…


- Micha Elsner, Cory Shain, and members of the OSU Clippers group for comments and suggestions

- Gabriel Stanovsky and Rachel Rudinger for sharing the data and code

- National Science Foundation for supporting part of this work

# Appendix

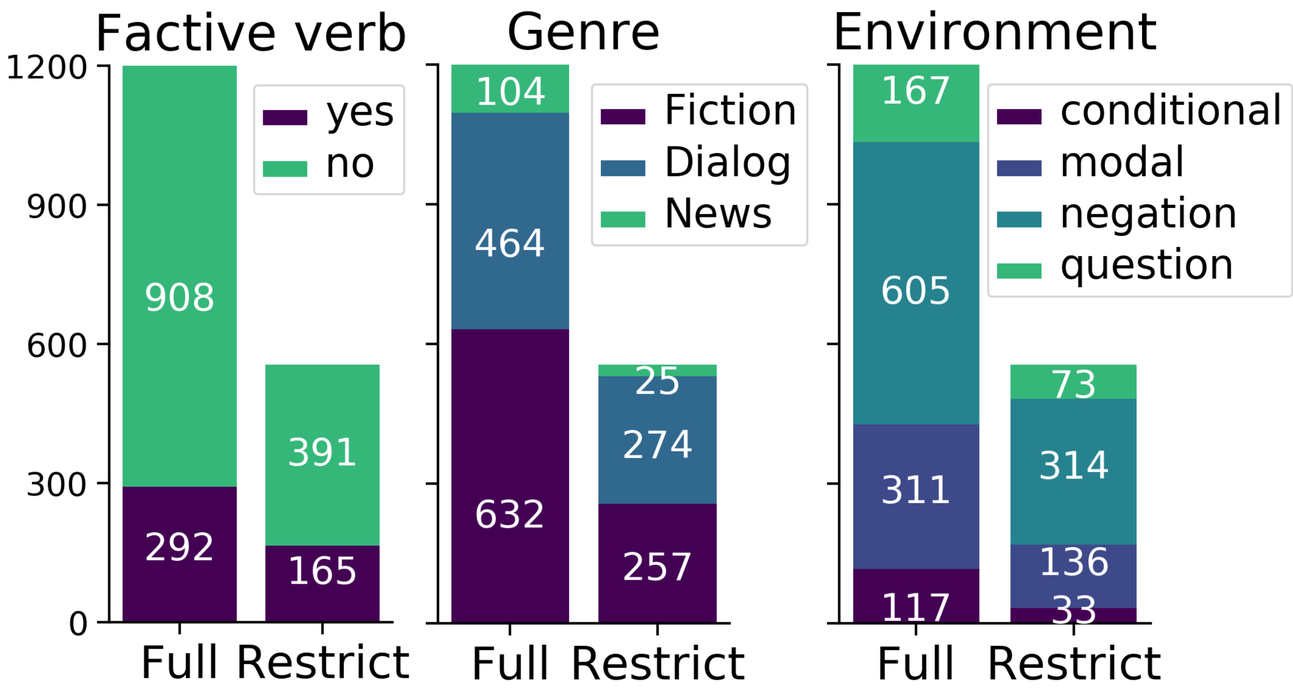
# CommitmentBank

- 1200 naturally-occurring English discourses
- 3 genres
- 4 entailment-canceling environments
- 48 clause-embedding verbs
- 8+ annotations/item
- 0.53 Krippendorff’s alpha

de Marneffe et al (2019)

# CommitmentBank restricted set

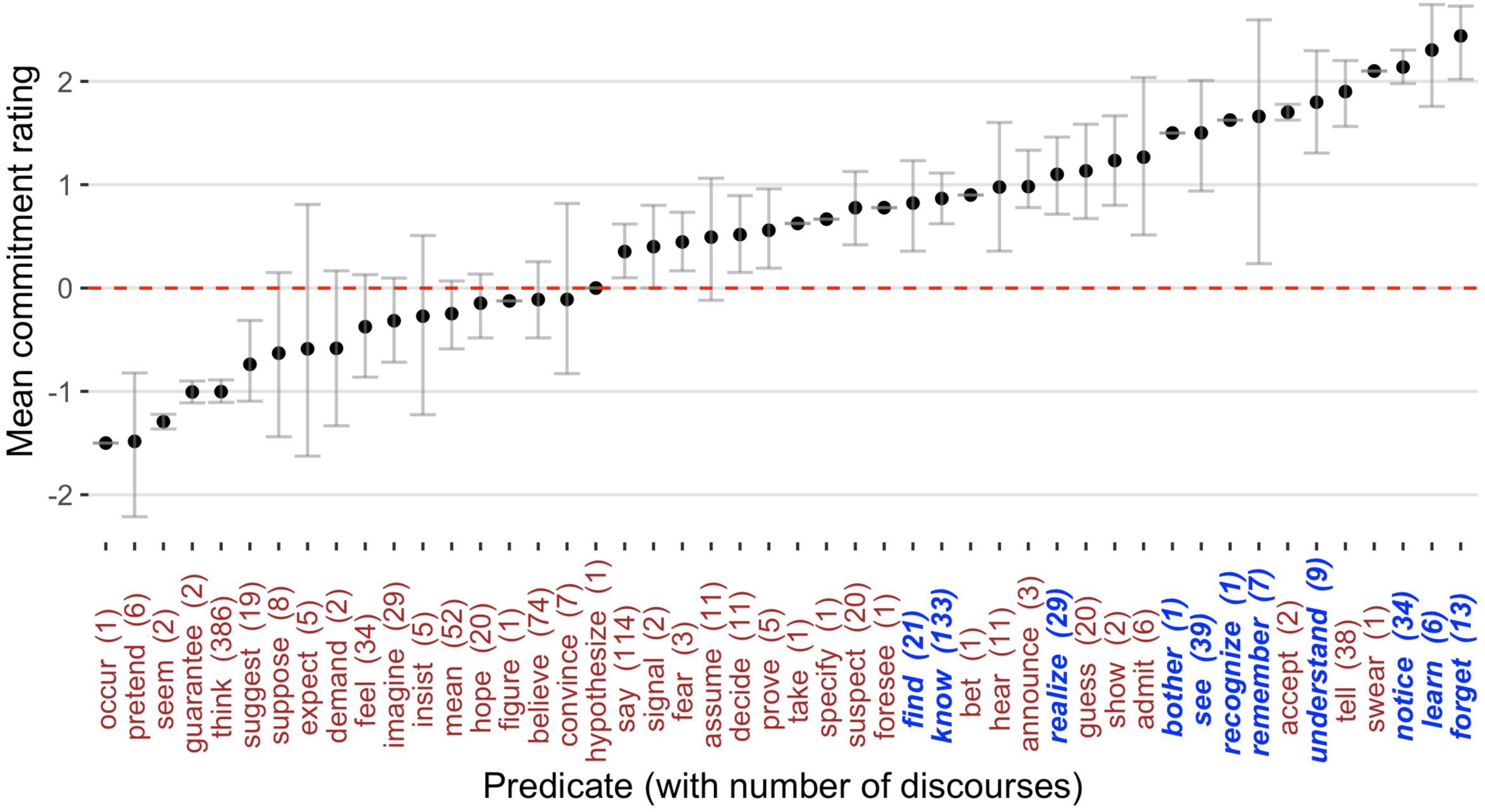
- 556 items
- 3 genres
- 37 clause-embedding verbs
- 8+ annotations/item
- 0.74 Krippendorff's alpha




**Proportions of items with each feature.**

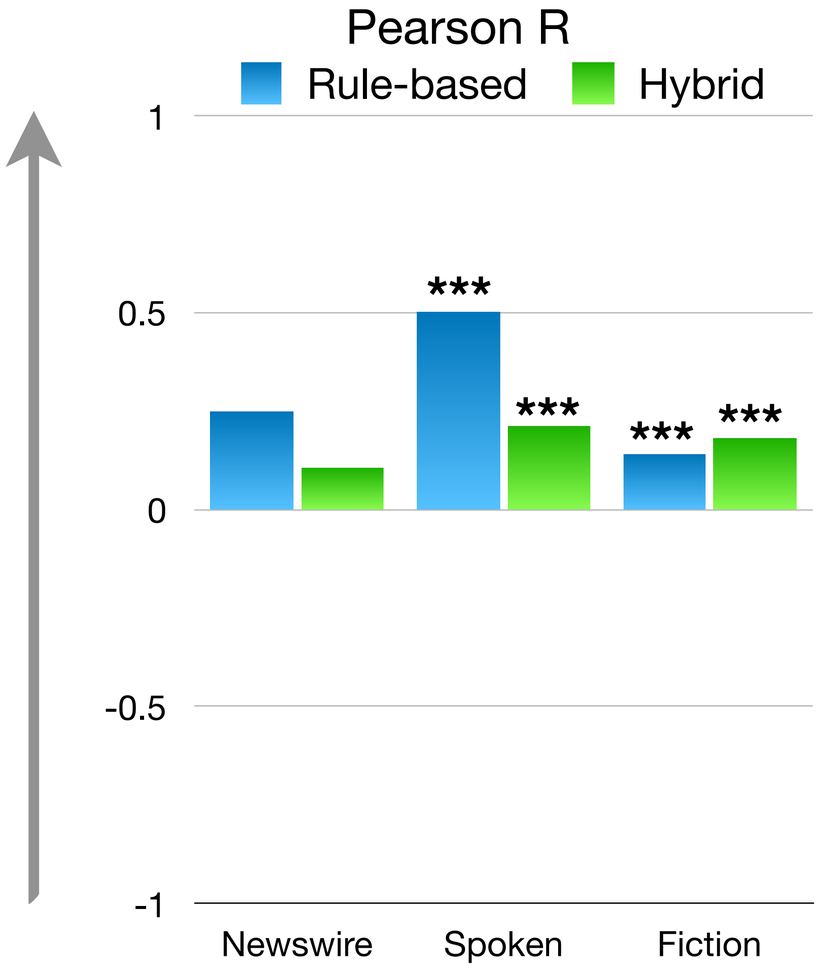
| | # Predicate | SoA | |
|---|---|---|---|
| | | $r$ | MAE |
| FactBank | 9761 | 0.86 | 0.31 |
| MEANTIME | 1395 | 0.61 | 0.23 |
| UW | 13644 | 0.75 | 0.42 |
| UDS | 27289 | 0.77 | 0.96 |

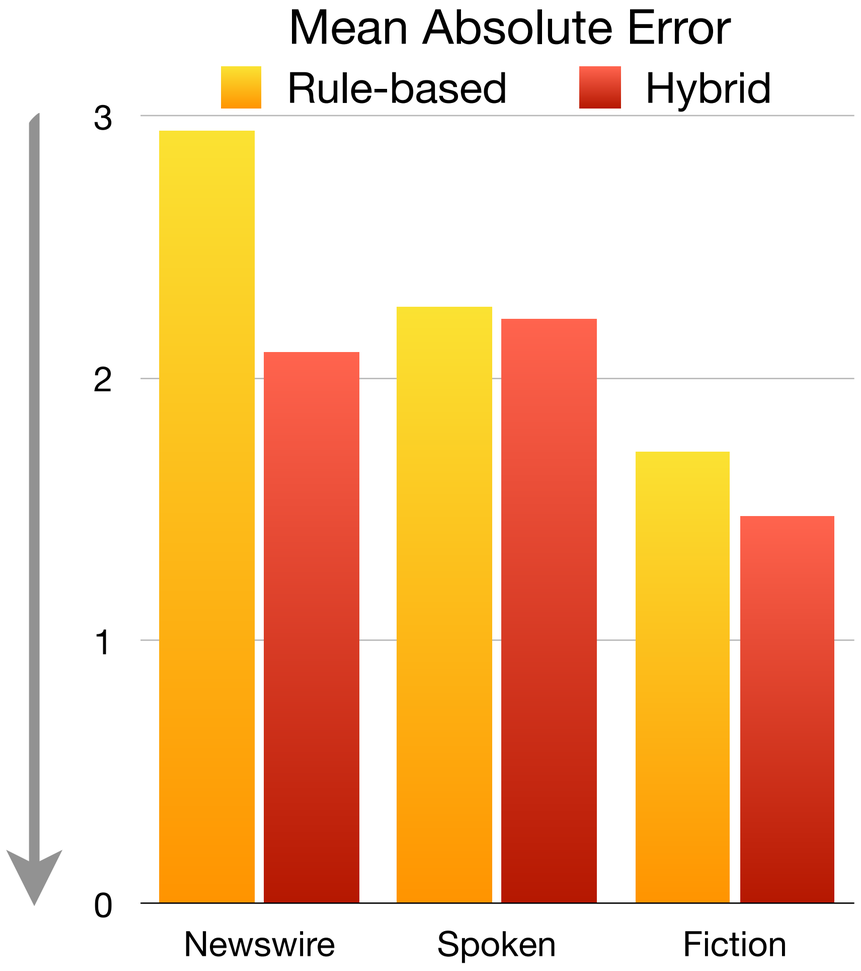
Table 2: The number of annotated predicates in each dataset, and previous state-of-the-art performance. The score on UW with MAE was obtained by Stanovsky et al. (2017), while the other scores were obtained by Rudinger et al. (2018).

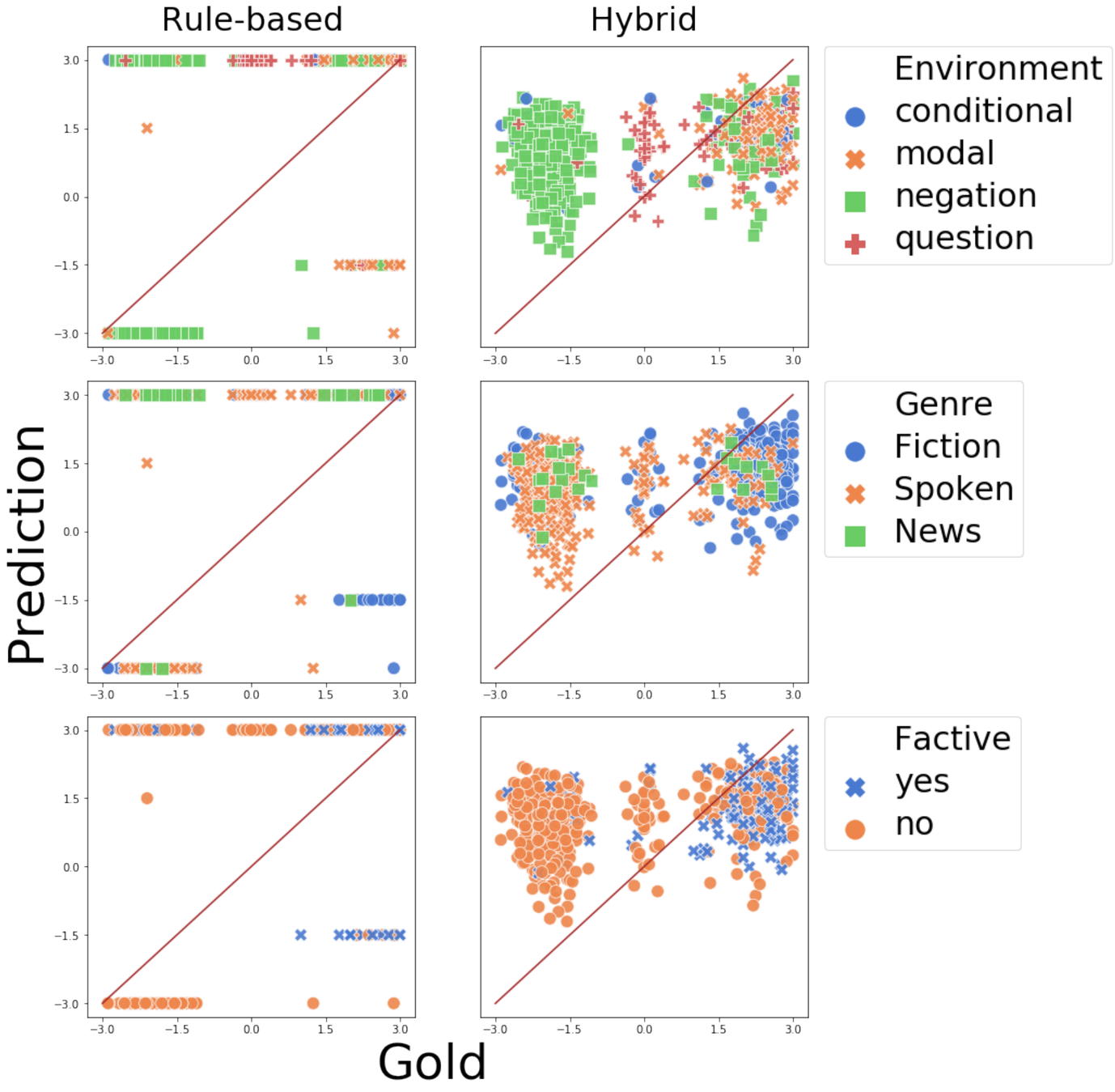

Mean commitment rating
2
1
0
-1
-2
occur (1)
pretend (6)
seem (2)
guarantee (2)
think (386)
suggest (19)
suppose (8)
expect (5)
demand (2)
feel (34)
imagine (29)
insist (5)
mean (52)
hope (20)
figure (1)
believe (74)
convince (7)
hypothesize (1)
say (114)
signal (2)
fear (3)
assume (11)
decide (11)
prove (5)
take (1)
specify (1)
suspect (20)
foresee (1)
find (21)
know (133)
bet (1)
hear (11)
announce (3)
realize (29)
guess (20)
show (2)
admit (6)
bother (1)
see (39)
recognize (1)
remember (7)
accept (2)
understand (9)
tell (38)
swear (1)
notice (34)
learn (6)
forget (13)
Predicate (with number of discourses)

# Genre

## Spoken language the best, newswire the worst




Rule-based
Hybrid
Prediction
Gold
Environment
conditional
modal
negation
question
Genre
Fiction
Spoken
News
Factive
yes
no
3.0
1.5
0.0
−1.5
−3.0