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

Nanjiang Jiang and Marie-Catherine de Marneffe
Department of Linguistics
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?

Do you know that Florence is packed with visitors?

Clause-embedding verb

Entailment-canceling environment

Question Negation Modal Conditional

Embedded clause

Commitment
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.
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: Mary is certain that it is true
- 2: Mary is not certain whether it is true or false
- 1: Mary is uncertain whether it is true or false
- 0: Mary is uncertain whether it is true or false
- -1: Mary is certain that it is false
- -2: Mary is not certain whether it is true or false
- -3: 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
Kim failed to remember to have 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 ∈ [-3, 3]

<table>
<thead>
<tr>
<th>Metric</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson R Correlation</td>
<td>[-1, 1]</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>[0, 3]</td>
</tr>
<tr>
<td>Variability</td>
<td></td>
</tr>
<tr>
<td>Absolute fit</td>
<td></td>
</tr>
</tbody>
</table>

More informative for biased datasets
Results

Performance is not great overall

Pearson R

Mean Absolute Error
Analysis
Focus on the best performing models

Pearson R

Mean Absolute Error
Entailment-canceling Environment

Negation the best, conditional the worst

Pearson R

<table>
<thead>
<tr>
<th></th>
<th>Rule-based</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negation</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean Absolute Error

<table>
<thead>
<tr>
<th></th>
<th>Rule-based</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Modal</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Negation</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>1.2</td>
<td></td>
</tr>
</tbody>
</table>
Factivity

Good absolute fit but poor correlation for factive verbs

Pearson R

Factive  Nonfactive  Rule-based  Hybrid

Mean Absolute Error

Factive  Nonfactive  Rule-based  Hybrid

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

<table>
<thead>
<tr>
<th>Committed to ( p )</th>
<th>Not committed</th>
<th>Committed to ( \neg p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>◎ 3</td>
<td>◎ 2</td>
<td>◎ 1</td>
</tr>
<tr>
<td>Mary is certain that it is true</td>
<td></td>
<td></td>
</tr>
<tr>
<td>◎ 0</td>
<td>◎ -1</td>
<td>◎ -2</td>
</tr>
<tr>
<td>Mary is not certain whether it is true or false</td>
<td></td>
<td></td>
</tr>
<tr>
<td>◎ -3</td>
<td></td>
<td>◎ -2</td>
</tr>
<tr>
<td>Mary is certain that it is false</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mary is certain that it is true
Mary is not certain whether it is true or false
Mary is certain that it is false

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.
<table>
<thead>
<tr>
<th># Predicate</th>
<th>$r$</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FactBank</td>
<td>0.86</td>
<td>0.31</td>
</tr>
<tr>
<td>MEANTIME</td>
<td>0.61</td>
<td>0.23</td>
</tr>
<tr>
<td>UW</td>
<td>0.75</td>
<td>0.42</td>
</tr>
<tr>
<td>UDS</td>
<td>0.77</td>
<td>0.96</td>
</tr>
</tbody>
</table>

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).
Genre
Spoken language the best, newswire the worst

Pearson R

- Rule-based
- Hybrid

Mean Absolute Error

- Rule-based
- Hybrid

Newswire  Spoken  Fiction

Pearson R

- ***

Mean Absolute Error

- Newswire
- Spoken
- Fiction
Gold

Rule-based

Hybrid

Environment
- conditional
- modal
- negation
- question

Genre
- Fiction
- Spoken
- News

Factive
- yes
- no

Prediction

Gold