Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment

Nanjiang Jiang and **Marie-Catherine de Marneffe**
Department of Linguistics
The Ohio State University
{jiang.1879, demarneffe.1}@osu.edu

Abstract

When a speaker, Mary, asks *Do you know that Florence is packed with visitors?*, we take her to believe that Florence is packed with visitors, but not if she asks *Do you think that Florence is packed with visitors?* Inferring speaker commitment (aka event factuality) is crucial for information extraction and question answering. Here we explore the hypothesis that linguistic deficits drive the error patterns of speaker commitment models by analyzing the linguistic correlates of model errors on a challenging naturalistic dataset. We evaluate two state-of-the-art speaker commitment models on the CommitmentBank, an English dataset of naturally occurring discourses. The CommitmentBank is annotated with speaker commitment towards the content of the complement (*Florence is packed with visitors* in our example) of clause-embedding verbs (*know*, *think*) under four entailment-canceling environments. We found that a linguistically-informed model outperforms a LSTM-based one, suggesting that linguistic knowledge is needed to capture such challenging naturalistic data. A breakdown of items by linguistic features reveals asymmetrical error patterns: while the models achieve good performance on some classes (e.g., negation), they fail to generalize to the diverse linguistic constructions (e.g., conditionals) in natural language, highlighting directions for improvement.

1 Introduction

Prediction of speaker commitment[1] is the task of determining to what extent the speaker is committed to an event in a sentence as actual, non-actual, or uncertain. This matters for downstream NLP applications, such as information extraction or question answering: for instance, we should extract from example (1) in Table 1 that the speaker could wish someone dead, but from (3) that people should *not* be allowed to carry guns in their vehicles, even though both events are embedded under *believe* and negation.

There has been work on factors leading to speaker commitment in theoretical linguistics (i.a., Karttunen (1971); Simons et al. (2010)) and computational linguistics (i.a., Diab et al. (2009); Saurí and Pustejovsky (2012); Prabhakaran et al. (2015)), but mostly on constructed or newswire examples, which may simplify the task by failing to reflect the lexical and syntactic diversity of naturally occurring utterances. de Marneffe et al. (2019) introduced the CommitmentBank, a dataset of naturally occurring sentences annotated with speaker commitment towards the content of complements of clause-embedding verbs under canceling-entailment environments (negation, modal, question and conditional), to study the linguistic correlates of speaker commitment. In this paper, we use it to evaluate two state-of-the-art (SoA) models of speaker commitment: Stanovsky et al. (2017) and Rudinger et al. (2018). The CommitmentBank, restricted to specific linguistic constructions, is a good test case. It allows us to evaluate whether current speaker commitment models achieve robust language understanding, by analyzing their performance on specific challenging linguistic constructions.

2 The CommitmentBank corpus

The CommitmentBank[2] consists of 1,200 naturally occurring items involving clause-embedding verbs under four entailment-canceling environments (negations, modals, questions, condition-

[1] Previous work uses event factuality, verdicality, or committed belief; the terms refer to the same linguistic phenomenon, perhaps with different emphasis.

[2] The data is available at https://github.com/mcdm/CommitmentBank

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4208–4213
Florence, Italy, July 28 - August 2, 2019.

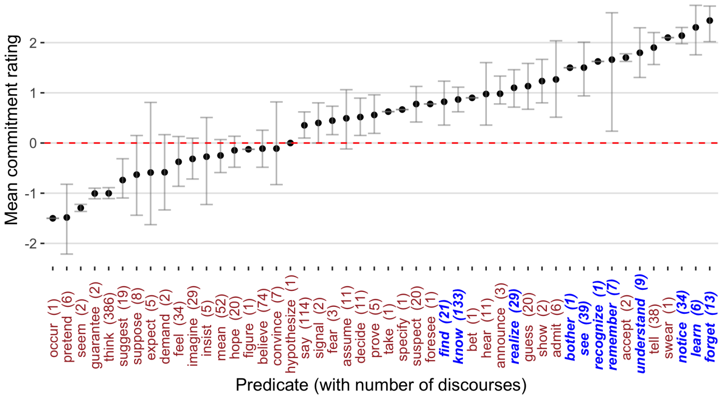

Figure 1: Mean commitment scores in all of the CommitmentBank. Italicized verbs are factive, plain nonfactive.

(1)	**Context**	The answer is no, no no. Not now, not ever.
	Target	I *never* **believed** $\underbrace{\text{I could wish anyone dead}}_{\text{Gold:1.56, Rule-based:3.0, Hybrid: 0.50}}$ but last night changed all that.
(2)	**Context**	Revenue is estimated at \$18.6 million. The maker of document image processing equipment said the state procurement division had declared FileNet in default on its contract with the secretary of state uniform commercial code division.
	Target	FileNet said it *doesn't* **believe** $\underbrace{\text{the state has a valid basis of default and is reviewing its legal rights under the contract}}_{\text{Gold: -0.47, Rule-based: 3.0, Hybrid: 1.08}}$, but said it can't predict the outcome of the dispute.
(3)	**Context**	**A**: Yeah, that's crazy. **B**: and then you come here in the Dallas area, um,
	Target	I *don't* **believe** that $\underbrace{\text{people should be allowed to carry guns in their vehicled}}_{\text{Gold: -2.64, Rule-based: 3.0, Hybrid: 1.40}}$.

Table 1: Examples from the CommitmentBank, with gold scores and predictions from rule-based and hybrid models. Embedding verbs in bold, entailment-canceling environments italicized. The gold score is the mean annotators' speaker commitment judgments towards the content of the complement.

als). Three genres are represented: newswire from the Wall Street Journal (WSJ), fiction from the British National Corpus, and dialog from Switchboard. Each item consists of up to two context sentences and one target sentence, as shown in Table 1. For each item, speaker commitment judgments were gathered on Mechanical Turk from at least eight native English speakers. Participants judged whether or not the speaker is certain that the content of the complement in the target sentence is true, using a Likert scale labeled at 3 points (+3/speaker is certain that the complement is true, 0/speaker is not certain whether it is true or false, -3/speaker is certain that it is false). We took the mean annotations of each item as gold score of speaker commitment. Figure 1 shows the mean annotations per embedding verb.

Restricted set We identified a subset of the CommitmentBank that displays high agreement among annotators. We divided the range of integer ratings $[-3, 3]$ into three sub-ranges: $[1, 3]$ where the speaker is committed to the complement p, 0 where the speaker is uncommitted towards p, $[-3, -1]$ where the speaker is committed to $\neg p$. We selected the items for which at least 80% of the annotations fall into the same sub-range. This gives 556 items, with 37 clause-embedding verbs. Figure 2 shows that the proportion of items with different linguistic features in the restricted set is similar to the proportion in the full set, suggesting that the restricted set is representative of the original data. The full CommitmentBank has a Krippendorff's α of 0.53, while α is 0.74 on the restricted set.

3 Models of speaker commitment

We evaluate the performance of two speaker commitment models on the CommitmentBank: a rule-based model (Stanovsky et al., 2017) and a neural-based one (Rudinger et al., 2018).

Rule-based model Stanovsky et al. (2017) proposed a rule-based model based on a deterministic algorithm based on TruthTeller (Lotan et al., 2013), which uses a top-down approach on a de-

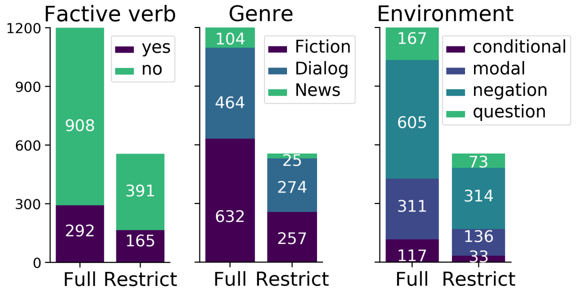

Figure 2: Number of items with different features in the full and restricted sets of the CommitmentBank.

pendency tree and predicts speaker commitment score in $[-3, 3]$ according to the implicative signatures (Karttunen, 2012) of the predicates, and whether the predicates are under the scope of negation and uncertainty modifiers. For example, *refuse* p entails $\neg p$, so the factuality of its complement p gets flipped if encountered.

Neural-based model Rudinger et al. (2018) introduced three neural models for speaker commitment: a linear biLSTM, a dependency tree biLSTM, a hybrid model that ensembles the two. Rudinger et al. (2018) also proposed a multi-task training scheme in which a model is trained on four factuality datasets: FactBank (Saurí and Pustejovsky, 2009), UW (Lee et al., 2015), MEANTIME (Minard et al., 2016) and UDS (Rudinger et al., 2018), all with annotations on a $[-3, 3]$ scale. Each dataset has shared biLSTM weights but specific regression parameters.

Reference datasets The FactBank, UW, and MEANTIME datasets all consist of sentences from news articles. Each event in FactBank was annotated by 2 annotators, with 0.81 Cohen's κ. UW has 5 annotations for each event, and MEANTIME has 6. UDS contains sentences from the English Web Treebank (Bies et al., 2012), which contains weblogs, newsgroups, emails, reviews, and question-answers. It has 2 annotations for each predicate, with 0.66 Cohen's κ. All four datasets have annotations biased towards +3, because (1) they are newswire-heavy with sentences describing known factual events, and (2) most annotations are for main-clause predicates instead of predicates in an embedded clause.

Table 2 gives the number of predicates in each dataset and state-of-the-art results obtained. Two metrics were reported for both models: mean absolute error (MAE), measuring the absolute fit, and Pearson's r correlation, measuring how well the model captures variability in the data. Pearson's r is considered more informative than MAE because the reference sets are biased towards +3.

	# Predicate	SoA	
		r	MAE
FactBank	9,761	0.86	0.31
MEANTIME	1,395	0.61	0.23
UW	13,644	0.75	0.42
UDS	27,289	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).

4 Evaluation

We evaluated the models of Stanovsky et al. (2017) and Rudinger et al. (2018) on the CommitmentBank. We used Stanovsky et al. (2017)'s rule-based annotator[3] to get commitment ratings for the embedded predicates of the target sentences. Following Rudinger et al. (2018), we trained the linear, tree, and hybrid biLSTM models using the multi-task training scheme on the four factuality datasets they used, which produced four predictions. Following White et al. (2018), we used cross-validated ridge regression to predict a final score using the four predictions.

We include a majority baseline "All -2.0" (always predicting -2.0, since -2.0 is the most frequent answer in the full and restricted CommitmentBank). The results are shown in Figure 3. The rule-based model outperforms the biLSTM models on the full set, but overall both SoA models do not perform very well on the CommitmentBank. As shown in Figure 3, the CommitmentBank is substantially more challenging for these models than the reference datasets, with lower correlation and higher absolute error rates than were obtained for any of these other datasets.

5 Analysis

Focusing on the restricted set, we perform detailed error analysis of the outputs of the rule-based and hybrid biLSTM models, which achieved the best

[3] https://github.com/gabrielStanovsky/unified-factuality

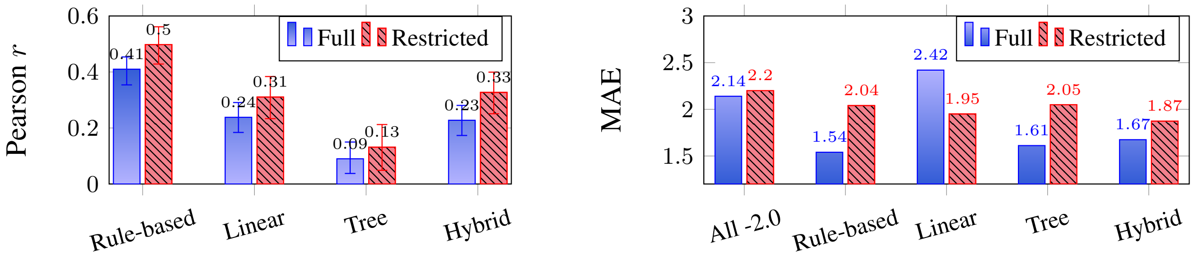

Figure 3: Pearson r correlation and Mean Absolute Error (MAE) on All -2.0 baseline, Rule-based annotator (Stanovsky et al., 2017), and three biLSTM models in Rudinger et al. (2018). Pearson r is undefined for All -2.0. All correlations are statistically significant ($p < 0.05$).

Feature	Value	r		MAE	
		Rule	Hybr.	Rule	Hybr.
Embedding	Cond.	NA	0.02	2.08	1.50
	Modal	-0.01	0.21 †	**1.37**	**1.08**
	Negation	**0.45** †	0.22 †	2.26	2.40
	Question	-0.22	**0.29** †	2.35	1.25
Genre	Fiction	0.25 †	0.18 †	**1.72**	**1.47**
	Dialog	**0.50** †	**0.21** †	2.27	2.23
	News	0.14	0.10	2.94	2.10
Factive	Yes	-0.14	0.23 †	**1.58**	**1.20**
	No	**0.49** †	**0.25** †	2.23	2.16
NegRaising		0.04	-0.07	1.91	2.77

Table 3: Performance and number of items per feature. The scores in bold indicate the classes on which each model has the best performance (with respect to both metrics). † marks statistical significance of Pearson's correlation ($p < 0.05$).

correlation. Table 3 shows performance for the following linguistic features, and Figure 4 shows scatterplots of gold judgments vs. predictions.

Embedding environment The rule-based model can only capture inferences involving negation ($r = 0.45$), while the hybrid model performs more consistently across negation, modal, and question ($r \sim 0.25$). Both models cannot handle inferences with conditionals.

The model's performance on the negation items also vary with respect to genre: the rule-based model has significant correlations for fiction ($r = 0.45$) and dialog ($r = 0.32$), while the hybrid model has correlations between 0.05 and 0.2 for all three genres, none reaching significance. About 40% of the modal and question items involve factive verbs, therefore the performance of these environments also correlate with the models' performance on factive verbs (elaborated on below).

Genre Both models achieve the best correlation on dialog (Switchboard), and the worst on newswire (WSJ). The poor performance on WSJ might be due to its scores in CommitmentBank being more widespread (reflected in Figure 4) than annotations in the reference datasets (e.g., MEANTIME), which tend to be biased towards +3. The good performance of the rule-based model on dialog could be due to the fact that 70% of the items in dialog are in a negation environment with a nonfactive verb.

Factive embedding verb Lexicalist theories (i.a., Karttunen 1973; Heim 1983) predict that complements of factive verbs are commitments of the speaker. This tendency is reflected in Figures 1 and 4 where most sentences with factives have higher mean commitment scores. Both models get better MAE on factives, but better correlation on nonfactives. The improved MAE of the rule-based model might be due to its use of factive/implicative signatures. However, the poor correlations suggest that neither model can robustly capture the variability in inference which exists in sentences involving factive/nonfactive verbs (see i.a. Beaver 2010; de Marneffe et al. 2019).

Neg-raising Within sentences with negation, we examine the models' performance on sentences with "neg-raising" reading, where a negation in the matrix clause (*not* {*think/believe*} p) is interpreted as negating the complement clause (*think/believe not* p), as in example (3) in Table 1 where we understand the speaker to be committed to *people should* ***not*** *be allowed to carry guns in their vehicles*. We identify "neg-raising" items as items with a negation embedding environment, *think* or *believe* verb, and a negative commitment score. There is almost no correlation between both

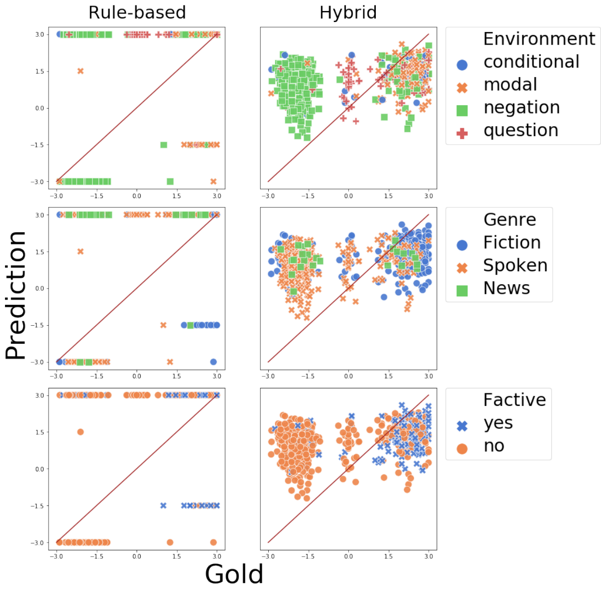

Figure 4: Gold scores vs. model prediction. Each point is a sentence. Lines show perfect predictions.

models' predictions and gold judgments (Table 3), suggesting that the models are not able to capture neg-raising inferences.

Model behavior Figure 4 shows that the hybrid model predictions are mostly positive, whereas the rule-based model predictions are clustered at −3 and +3. This suggests that the rule-based model cannot capture the gradience present in commitment judgments, while the hybrid model struggles to recognize negative commitments.

To better interpret the models' outputs, we evaluate them in a classification setting. We use Gaussian mixture models to obtain three clusters for the mean gold scores and the predictions of both models. We assign the cluster with the highest mean to +: speaker is certain that the complement is true, the one with the lowest mean to −: speaker is certain that it is false, and the remaining one to ○: speaker is not certain about its truth. We report precision, recall and F1 in Table 4. The rule-based model predicts + by default unless it has clear evidence (e.g., negation) for negative commitment. This behavior is reflected in the high precision for −. Both models perform well on + and −, but neither is able to identify no commitment (○).

6 Conclusion

Our evaluation of two SoA models for speaker commitment on the CommitmentBank shows that the models perform better on sentences with negation, and with nonfactive embedding verbs. However, they are not able to generalize to other linguistic environments such as conditional, modal, and neg-raising, which display inference patterns that are important for information extraction. Both models are able to identify the polarity of commitment, but cannot capture its gradience. The rule-based model, outperforming the biLSTM models on the full CommitmentBank, shows that a linguistically-informed model scales more successfully to challenging naturalistic data.

In the long run, to perform robust language understanding, models will need to incorporate more linguistic foreknowledge and be able to generalize to a wider range of linguistic constructions.

	Precision		Recall		F1		Count
	Rule	Hybr.	Rule	Hybr.	Rule	Hybr.	
+	0.58	0.64	0.91	0.51	0.71	0.56	251
−	0.99	0.67	0.55	0.20	0.70	0.31	268
○	0.00	0.06	0.00	0.46	0	0.11	37
Total	0.74	0.61	0.67	0.35	0.66	0.41	556

Table 4: Classification performance of the models.

Acknowledgment

We thank the anonymous reviewers for their valuable feedback. We also thank Micha Elsner, Cory Shain, and the Clippers group at The Ohio State University for their helpful discussions and suggestions. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1464252.

References

David Beaver. 2010. Have you noticed that your belly button lint colour is related to the colour of your clothing? In Rainer Bäuerle, Uwe Reyle, and Thomas Ede Zimmermann, editors, *Presuppositions and Discourse: Essays Offered to Hans Kamp*, pages 65–99. Leiden, The Netherlands: Brill.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73.

Irene Heim. 1983. On the projection problem for presuppositions. In *West Coast Conference on Formal Linguistics (WCCFL) 2*, pages 114–125.

Lauri Karttunen. 1971. Some observations on factivity. *Papers in Linguistics*, 4:55–69.

Lauri Karttunen. 1973. Presuppositions and compound sentences. *Linguistic Inquiry*, 4(2):169–193.

Lauri Karttunen. 2012. Simple and phrasal implicatives. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 124–131.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. TruthTeller: Annotating predicate truth. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–757.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentBank: Investigating projection in naturally occurring discourse. In *Sinn und Bedeutung 23*.

Anne-Lyse Myriam Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the newsreader multilingual event and time corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4417–4422.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 731–744.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Proceedings of Semantics and Linguistic Theory 20*. CLC Publications.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724.