He Thinks He Knows Better than the Doctors: BERT for Event Factuality Fails on Pragmatics

Nanjiang Jiang
Department of Linguistics
The Ohio State University, USA
jiang.1879@osu.edu

Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University, USA
demarneffe.1@osu.edu

Abstract

We investigate how well BERT performs on predicting factuality in several existing English datasets, encompassing various linguistic constructions. Although BERT obtains a strong performance on most datasets, it does so by exploiting common surface patterns that correlate with certain factuality labels, and it fails on instances where pragmatic reasoning is necessary. Contrary to what the high performance suggests, we are still far from having a robust system for factuality prediction.

1 Introduction

Predicting event factuality[1] is the task of identifying to what extent an event mentioned in a sentence is presented by the author as factual. It is a complex semantic and pragmatic phenomenon: in *John thinks he knows better than the doctors*, we infer that John probably doesn't know better than the doctors. Event factuality inference is prevalent in human communication and matters for tasks that depend on natural language understanding, such as information extraction. For instance, in the FactBank example (Saurí and Pustejovsky, 2009) in Table 1, an information extraction system should extract *people are stranded without food* but not *helicopters located people stranded without food*.

The current state-of-the-art model for factuality prediction on English is the work of Pouran Ben Veyseh et al. (2019), obtaining the best performance on four factuality datasets: FactBank, MEANTIME (Minard et al., 2016), UW (Lee et al., 2015), and UDS-IH2 (Rudinger et al., 2018). Traditionally, event factuality is thought to be triggered by fixed properties of lexical items. The Rule-based model of Stanovsky et al. (2017) took such an approach: They used lexical rules and dependency trees to determine whether an event in a sentence is factual, based on the properties of the lexical items that embed the event in question. Rudinger et al. (2018) proposed the first end-to-end model for factuality with LSTMs. Pouran Ben Veyseh et al. (2019) used BERT representations with a graph convolutional network and obtained a large improvement over Rudinger et al. (2018) and over Stanovsky et al.'s (2017) Rule-based model (except for one metric on the UW dataset).

However, it is not clear what these end-to-end models learn and what features are encoded in their representations. In particular, they do not seem capable of generalizing to events embedded under certain linguistic constructions. White et al. (2018) showed that the Rudinger et al. (2018) models exhibit systematic errors on MegaVeridicality, which contains factuality inferences purely triggered by the semantics of clause-embedding verbs in specific syntactic contexts. Jiang and de Marneffe (2019a) showed that Stanovsky et al.'s and Rudinger et al.'s models fail to perform well on the CommitmentBank (de Marneffe et al., 2019), which contains events under clause-embedding verbs in an entailment-canceling environment (negation, question, modal, or antecedent of conditional).

In this paper, we investigate how well BERT, using a standard fine-tuning approach,[2] performs on seven factuality datasets, including those focusing on embedded events that have been shown to be challenging (White et al., 2018 and Jiang and de Marneffe 2019a). The application of BERT to datasets focusing on embedded events has been limited to the setup of natural language inference (NLI) (Poliak et al., 2018; Jiang and de Marneffe, 2019b; Ross and Pavlick, 2019). In the NLI setup,

[1] The terms *veridicality* and *speaker commitment* refer to the same underlying linguistic phenomenon.

[2] We only augment BERT with a task-specific layer, instead of proposing a new task-specific model as in Pouran Ben Veyseh et al. (2019).

Transactions of the Association for Computational Linguistics, vol. 9, pp. 1081–1097, 2021. https://doi.org/10.1162/tacl_a_00414
Action Editor: Benjamin Van Durme. Submission batch: 1/2021; Revision batch: 5/2021; Published 10/2021.

MegaVeridicality	Someone was **misinformed** that something happened$^{-2.7}$.
CB	Hazel had not felt so much bewildered since Blackberry had talked about the raft beside the Enborne. Obviously, the stones could not possibly be anything to do with El-ahrairah. It seemed to him that Strawberry *might* as well have **said** that his tail was$^{-1.33}$ an oak tree.
RP	The man **managed** to stay^{3} on his horse. / The man did *not* **manage** to stay$^{-2.5}$ on his horse.
FactBank	Helicopters are flying$^{3.0}$ over northern New York today **trying**$^{3.0}$ to locate^{0} people stranded$^{3.0}$ without food, heat or medicine.
MEANTIME	Alongside both announcements$^{3.0}$, Jobs also announced$^{3.0}$ a new iCloud service to sync^{0} data among all devices.
UW	Those plates may have come$^{1.4}$ from a machine shop in north Carolina, where a friend of Rudolph worked$^{3.0}$.
UDS-IH2	DPA: Iraqi authorities **announced**$^{2.25}$ that they had busted$^{2.625}$ up 3 terrorist cells operating$^{2.625}$ in Baghdad.

Table 1: Example items from each dataset. The annotated event predicates are underlined with their factuality annotations in superscript. For the datasets focusing on embedded events (first group), the clause-embedding verbs are in bold and the entailment-canceling environments (if any) are *slanted*.

an item is a premise-hypothesis pair, with a categorical label for whether the event described in the hypothesis can be inferred by the premise. The categorical labels are obtained by discretizing the original real-valued annotations. For example, given the premise *the man managed to stay on his horse* (RP example in Table 1) and the hypothesis *the man stayed on his horse*, a model should predict that the hypothesis can be inferred from the premise. In the factuality setup, an item contains a sentence with one or more spans corresponding to events, with real-valued annotations for the factuality of the event. By adopting the event factuality setup, we study whether models can predict not only the polarity but also the gradience in factuality judgments (which is removed in the NLI-style discretized labels). Here, we provide an in-depth analysis to understand which kind of items BERT fares well on, and which kind it fails on. Our analysis shows that, while BERT can pick up on subtle surface patterns, it consistently fails on items where the surface patterns do not lead to the factuality labels frequently associated with the pattern, and for which pragmatic reasoning is necessary.

2 Event Factuality Datasets

Several event factuality datasets for English have been introduced, with examples from each shown in Table 1. These datasets differ with respect to some of the features that affect event factuality.

Embedded Events The datasets differ with respect to which events are annotated for factuality. The first category, including MegaVeridicality (White et al., 2018), CommitmentBank (CB), and Ross and Pavlick (2019) (RP), only contains sentences with clause-embedding verbs and factuality is annotated solely for the event described by the embedded clause. These datasets were used to study speaker commitment towards the embedded content, evaluating theories of lexical semantics (Kiparsky and Kiparsky, 1970; Karttunen, 1971a; Beaver, 2010, among others), and probing whether neural model representations contain lexical semantic information. In the datasets of the second category (FactBank, MEANTIME, UW, and UDS-IH2), events in both main clauses and embedded clauses (if any) are annotated. For instance, the example for UDS-IH2 in Table 1 has annotations for the main clause event *announced* and the embedded clause event *busted*, while the example for RP is annotated only for the embedded clause event *stay*, but not for the main clause event *managed*.

Genres The datasets also differ in genre: FactBank, MEANTIME, and UW are newswire data. Because newswire sentences tend to describe factual events, these datasets have annotations biased towards factual. UDS-IH2, an extension of White et al. (2016), comes from the English Web Treebank (Bies et al., 2012) containing weblogs, emails, and other web text. CB comes from three genres: newswire (Wall Street Journal), fiction (British National Corpus), and dialog (Switchboard). RP contains short sentences sampled from MultiNLI (Williams et al., 2018) from 10 different genres. MegaVeridicality contains artificially constructed ‘‘semantically bleached’’ sentences to remove confound of pragmatics and world-knowledge, and to collect baseline judgments of how much the verb by itself affects the factuality of the content of its complement in certain syntactic constructions.

Entailment-canceling Environments The three datasets in the first category differ with respect

to whether the clause-embedding verbs are under some entailment-canceling environment, such as negation. Under the framework of implicative signatures (Karttunen, 1971a; Nairn et al., 2006; Karttunen, 2012), a clause-embedding verb (in a certain syntactic frame—details later) has a lexical semantics (a signature) indicating whether the content of its complement is factual (+), nonfactual (−), or neutral (○, no indication of whether the event is factual or not). A verb signature has the form X/Y, where X is the factuality of the content of the clausal complement when the sentence has positive polarity (not embedded under any entailment-canceling environment), and Y is the factuality when the clause-embedding verb is under negation. In the RP example in Table 1, *manage to* has signature +/− which, in the positive polarity sentence *the man managed to stay on his horse*, predicts the embedded event *stay* to be factual (such intuition is corroborated by the +3 human annotation). Conversely, in the negative polarity sentence *the man did not manage to stay on his horse*, the − signature signals that *stay* is nonfactual (again corroborated by the −2.5 human annotation). For *manage to*, negation cancels the factuality of its embedded event.

While such a framework assumes that different entailment-canceling environments (negation, modal, question, and antecedent of conditional) have the same effects on the factuality of the content of the complement (Chierchia and McConnell-Ginet, 1990), there is evidence for varying effects of environments. Karttunen (1971b) points out that, while the content of complement of verbs such as *realize* and *discover* stays factual under negation (compare (1) and (2)), it does not under a question (3) or in the antecedent of a conditional (4).

(1) I **realized** that I had not told the truth.$^{+}$

(2) I *didn't* **realize** that I had not told the truth.$^{+}$

(3) *Did* you **realize** that you had not told the truth$^{\circ}$?

(4) *If* I **realize** later that I have not told the truth$^{\circ}$, I will confess it to everyone.

Smith and Hall (2014) provided experimental evidence that the content of the complement of *know* is perceived as more factual when *know* is under negation than when it is in the antecedent of a conditional.

In MegaVeridicality, each positive polarity sentence is paired with a negative polarity sentence where the clause-embedding verb is negated. Similarly in RP, for each naturally occurring sentence of positive polarity, a minimal pair negative polarity sentence was automatically generated. The verbs in CB appear in four entailment-canceling environments: negation, modal, question, and antecedent of conditional.

Frames Among the datasets in the first category, the clause-embedding verbs are under different syntactic contexts/frames, which also affect the factuality of their embedded events. For example, *forget* has signature +/+ in *forget that S*, but −/+ in *forget to VP*. That is, in *forget that S*, the content of the clausal complement *S* is factual in both *someone forgot that S* and *someone didn't forget that S*. In *forget to VP*, the content of the infinitival complement *VP* is factual in *someone didn't forget to VP*, but not in *someone forgot to VP*.

CB contains only *VERB that S* frames. RP contains both *VERB that S* and *VERB to VP* frames. MegaVeridicality exhibits nine frames, consisting of four argument structures and manipulations of active/passive voice and eventive/stative embedded VP: *VERB that S, was VERBed that S, VERB for NP to VP, VERB NP to VP-eventive, VERB NP to VP-stative, NP was VERBed to VP-eventive, NP was VERBed to VP-stative, VERB to VP-eventive, VERB to VP-stative*.

Annotation Scales The original FactBank and MEANTIME annotations are categorical values. We use Stanovsky et al.'s (2017) unified representations for FactBank and MEANTIME, which contain labels in the $[-3, 3]$ range derived from to the original categorical values in a rule-based manner. The original annotations of MegaVeridicality contain three categorical values *yes/maybe/no*, which we mapped to 3/0/−3, respectively. We then take the mean of the annotations for each item. The original annotations in RP are integers in $[-2, 2]$. We multiply each RP annotation by 1.5 to obtain labels in the same range as in the other datasets. The mean of the converted annotations is taken as the gold label for each item.

3 Linguistic Approaches to Factuality

Most work in NLP on event factuality has taken a lexicalist approach, tracing back factuality to

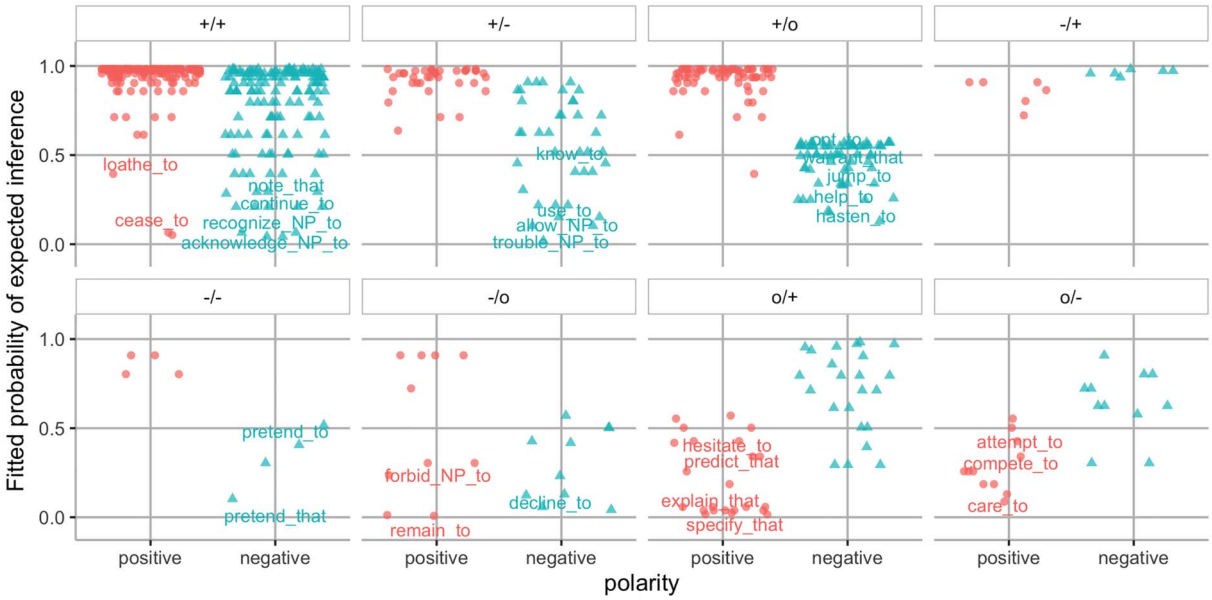

Figure 1: Fitted probabilities of true expected inference category predicted by the label of each item given by the ordered logistic regression model, organized by the signature and polarity. Some examples of verb-frame with mean probability less than 0.5 are labeled.

fixed properties of lexical items. Under such an approach, properties of the lexical patterns present in the sentence determine the factuality of the event, without taking into account contextual factors. We will refer to the inference calculated from lexical patterns only as *expected inference*. For instance, in (5), the expected inference for the event *had* embedded under *believe* is neutral. Indeed, because both true and false things can be believed, one should not infer from *A believes that S* that *S* is true (in other words, *believe* has as ○ / ○ signature), making *believe* a so-called ‘‘non-factive’’ verb by opposition to ‘‘factive’’ verbs (such as *know* or *realize*, which generally entail the truth of their complements both in positive polarity sentences (1) and in entailment-canceling environments (2), Kiparsky and Kiparsky [1970]). However, lexical theories neglect the pragmatic enrichment that is pervasive in human communication and fall short in predicting the correct inference in (5), where people judged the content of the complement to be true (as indicated by the annotation score of 2.38).

(5) Annabel *could* hardly **believe** that she had$^{2.38}$ a daughter about to go to university.

In FactBank, Saurí and Pustejovsky (2009) took a lexicalist approach, seeking to capture only the effect of lexical meaning and knowledge local to the annotated sentence: Annotators were linguistically trained and instructed to avoid using knowledge from the world or from the surrounding context of the sentence. However, it has been shown that such annotations do not always align with judgments from linguistically naive annotators. de Marneffe et al. (2012) and Lee et al. (2015) re-annotated part of FactBank with crowdworkers who were given minimal guidelines. They found that events embedded under report verbs (e.g., *say*), annotated as neutral in FactBank (since, similarly to *believe*, one can report both true and false things), are often annotated as factual by crowdworkers. Ross and Pavlick (2019) showed that their annotations also exhibit such a *veridicality bias*: Events are often perceived as factual/nonfactual, even when the expected inference specified by the signature is neutral. The reason behind this misalignment is commonly attributed to pragmatics: Crowdworkers use various contextual features to perform pragmatic reasoning that overrides the expected inference defined by lexical semantics. There has been theoretical linguistics work arguing that factuality is indeed tied to the discourse structure and not simply lexically controlled (among others, Simons et al., 2010).

Further, our analysis of MegaVeridicality shows that there is also some misalignment between the inference predicted by lexical semantics and the human annotations, even in cases without

not continue to signature: +/+, expected: +, observed: −
A particular person didn't continue to do$^{-0.33}$ a particular thing.
A particular person didn't continue to have$^{-1.5}$ a particular thing.
They did not continue to sit^{-3} in silence.
He did not continue to talk^{-3} about fish.
not pretend to signature: −/−, expected: −, observed: closer to ○
Someone didn't pretend to have$^{-1.2}$ a particular thing.
He did not pretend to aim$^{-0.5}$ at the girls.
{**add/warn**} **that** signature: ○/+, expected: ○, observed: +
Someone **added** that a particular thing happened$^{2.1}$.
Linda Degutis **added** that interventions have$^{2.5}$ to be monitored.
Someone **warned** that a particular thing happened$^{2.1}$.
It **warns** that Mayor Giuliani 's proposed pay freeze could destroy the NYPD 's new esprit de corps$^{2.5}$.
not {**decline/refuse**} **to** signature: −/○, expected: ○, observed: +
A particular person *didn't* **decline** to do$^{1.5}$ a particular thing.
We do *not* **decline** to sanction$^{2.5}$ such a result.
A particular person *didn't* **refuse** to do$^{2.1}$ a particular thing.
The commission did *not* **refuse** to interpret$^{2.0}$ it.

Table 2: Items with verbs that often behave differently from the signatures. The semantically bleached sentences are from MegaVeridicality, the others from RP. Gold labels are superscripted.

pragmatic factors. Recall that MegaVeridicality contains semantically bleached sentences where the only semantically loaded word is the embedding verb. We used ordered logistic regression to predict the expected inference category (+, ○, −) specified by the embedding verb signatures defined in Karttunen (2012) from the mean human annotations.[3] The coefficient for mean human annotations is 1.488 (with 0.097 standard error): Thus, overall, the expected inference aligns with the annotations.[4] However, there are cases where they diverge. Figure 1 shows the fitted probability of the true expected inference category for each item, organized by the signatures and polarity. If the expected inference was always aligning with the human judgments, the fitted probabilities would be close to 1 for all points. However, many points have low fitted probabilities, especially when the expected inference is ○ (e.g., negative polarity of +/○ and −/○, positive polarity of ○/+ and ○/−), showing that there is veridicality bias in MegaVeridicality, similar to RP. Table 2 gives concrete examples from MegaVeridicality and RP, for which the annotations often differ from the verb signatures: Events under *not refuse to* are systematically annotated as factual, instead of the expected neutral. The RP examples contain minimal content information (but the mismatch in these examples may involve pragmatic reasoning).

In any case, given that neural networks are function approximators, we hypothesize that BERT can learn these surface-level lexical patterns in the training data. But items where pragmatic reasoning overrides the lexical patterns would probably be challenging for the model.

4 Model and Experiment Setup

To analyze what BERT can learn, we use the seven factuality datasets in Table 1.

Data Preprocessing The annotations of CB and RP have been collected by asking annotators to rate the factuality of the content of the complement, which may contain other polarity and modality operators, whereas in FactBank annotators rated the factuality of the normalized complement, without polarity and modality operators. For example, the complement *anything should be done in the short term* contains the modal operator *should*, while the normalized complement would be *anything is done in the short term*. In MEANTIME, UW, and UDS-IH2, annotators rated the factuality of the event represented by a word in the original sentence, which has the effect of removing such operators. Therefore, to ensure a uniform interpretation of annotations between datasets, we semi-automatically identified items in CB and RP where the complement is not normalized,[5] for which we take the whole embedded clause to be the span for factuality prediction. Otherwise, we take the root of the embedded clause as the span.

We also excluded 236 items in RP where the event for which annotations were gathered cannot be represented by a single span from the sentence. For example, for *The Post Office is forbidden from ever attempting to close any office*, annotators were asked to rate the factuality of *the Post Office is forbidden from ever closing any office*. Simply taking the span *close any office* corresponds to the event of *the Post Office close any office*, but not to the event for which annotations are collected.

[3]The analysis is done on items with verb-frame combinations (and their passive counterparts) for which Karttunen (2012) gives a signature (i.e., 618 items from MegaVeridicality).

[4]The threshold for −|○ is −2.165 with SE 0.188. The threshold for ○|+ is 0.429 with SE 0.144.

[5]We automatically identified whether the complement contains a `neg` dependency relation, modal operators (*should, could, can, must, perhaps, might, maybe, may, shall, have to, would*), or adverbs, and manually verified the output.

	train	dev	test
MegaVeridicality	2,200	626	2,200
CommitmentBank	250	56	250
RP	1,100	308	1,100
FactBank	6,636	2,462	663
MEANTIME	1,012	195	188
UW	9,422	3,358	864
UDS-IH2	22,108	2,642	2,539

Table 3: Number of events in each dataset split.

Excluding Data with Low Agreement Annotation There are items in RP and CB that exhibit bimodal annotations. For instance, the sentence in RP *White ethnics have ceased to be the dominant force in urban life* received 3 annotation scores: -3/nonfactual, 1.5/between neutral and factual, and 3/factual. By taking the mean of such bimodal annotations, we end up with a label of 0.5/neutral, which is not representative of the judgments in the individual annotations. For RP (where each item received three annotations), we excluded 250 items where at least two annotations have different signs. For CB (where each item received at least 8 annotations), we follow Jiang and de Marneffe (2019a) by binning the responses into $[-3, -1]$, $[0]$, $[1, 3]$ and discarding items if less than 80% of the annotations fall in the same bin.

Data Splits We used the standard train/dev/test split for FactBank, MEANTIME, UW, and UDS-IH2. As indicated above, we only use the high agreement subset of CB with 556 items, with splits from Jiang and de Marneffe (2019b). We randomly split MegaVeridicality and RP with stratified sampling to keep the distributions of the clause-embedding verbs similar in each split. Table 3 gives the number of items in each split.

Model Architecture The task is to predict a scalar value in $[-3, 3]$ for each event described by a span in the input sentence. A sentence is fed into BERT and the final-layer representations for the event span are extracted. Because the spans have variable lengths, the SelfAttentiveSpanExtractor (Gardner et al., 2018) is used to weightedly combine the representations of multiple tokens and create a single vector for the original event span. The extracted span vectors are fed into a two-layer feed-forward network with tanh activation function to predict a single scalar value. Our architecture is similar to Rudinger et al.'s (2018) linear-biLSTM model, except that the input is encoded with BERT instead of bidirectional LSTM, and a span extractor is used. The model is trained with the smooth L1 loss.[6]

Evaluation Metrics Following previous work, we report mean absolute error (MAE), measuring absolute fit, and Pearson's r correlation, measuring how well models capture variability in the data. r is considered more informative since some datasets (MEANTIME in particular) are biased towards $+3$.

Model Training For all experiments, we fine-tuned BERT using the `bert_large_cased` model. Each model is fine-tuned with at most 20 epochs, with a learning rate of $1e-5$. Early stopping is used: Training stops if the difference between Pearson's r and MAE does not increase for more than 5 epochs. Most training runs last more than 10 epochs. The checkpoint with the highest difference between Pearson's r and MAE on the dev set is used for testing. We explored several training data combinations:

-Single: Train with each dataset individually;

-Shared: Treat all datasets as one;

-Multi: Datasets share the same BERT parameters while each has its own classifier parameters.

The Single and Shared setups may be combined with first fine-tuning BERT on MultiNLI, denoted by the superscript ^{M}. We tested on the test set of the respective datasets.

We also tested whether BERT improves on previous models on its ability to generalize to embedded events. The models in Rudinger et al. (2018) were trained on FactBank, MEANTIME, UW, and UDS-IH2 with shared encoder parameters and separate classifier parameters, and an ensemble of the four classifiers. To make a fair comparison, we followed Rudinger et al.'s setup by training BERT on FactBank, MEANTIME, UW, and UDS-IH2

[6]The code and data are available at `https://github.com/njjiang/factuality_bert`. The code is based on the toolkit `jiant v1` (Wang et al., 2019).

	R					Previous	MAE					Previous
	Shared	Shared^M	Single	Single^M	Multi	SotA	Shared	Shared^M	Single	Single^M	Multi	SotA
CB	0.865	0.869	0.831	0.878	0.89†		0.713	0.722	0.777	0.648	0.617†	
RP	0.806	0.813	0.867	0.867	0.87†		0.733	0.714	0.621	0.619	0.608†	
MegaVeridicality	0.876†	0.873	0.857	0.863	0.857		0.508	0.501†	0.531	0.523	0.533	
FactBank	0.836	0.845	0.914†	0.901	0.903	0.903	0.42	0.417	0.228†	0.241	0.236	0.31
MEANTIME	0.557	0.572†	0.503	0.513	0.491	0.702	0.333	0.338	0.355	0.345	0.319†	0.204
UW	0.776	0.787	0.868†	0.868	0.865	0.83	0.532	0.523	0.349†	0.351	0.351	0.42
UDS-IH2	0.845	0.843	0.855†	0.854	0.853	0.909	0.794	0.804	0.76†	0.763	0.766	0.726

Table 4: Performance on the test sets under different BERT training setups. The best score obtained by our models for each dataset under each metric is marked by †. The overall best scores are highlighted. Each score is the average from three runs with different random initialization. The previous state-of-the-art results are given when available. All come from Pouran Ben Veyseh et al. (2019), except the MAE score on UW, which comes from Stanovsky et al. (2017).

with one single set of parameters[7] and tested on MegaVeridicality and CommitmentBank.[8]

5 Results

Table 4 shows performance on the various test sets with the different training schemes. These models perform well and obtain the new state-of-the-art results on FactBank and UW, and comparable performance to the previous models on the other datasets (except for MEANTIME[9]). Comparing Shared vs. Shared^M and Single vs. Single^M, we see that transferring with MNLI helps all datasets on at least one metric, except for UDS-IH2 where MNLI-transfer hurts performance. The Multi and Single models obtain the best performance on almost all datasets other than MegaVeridicality and MEANTIME. The success of these models confirms the findings of Rudinger et al. (2018) that having dataset-specific parameters is necessary for optimal performance. Although this is expected, since each dataset has its own specific features, the resulting model captures data-specific quirks rather than generalizations about event factuality. This is problematic if one wants to deploy the system in downstream applications, since which dataset the input sentence will be more similar to is unknown a priori.

[7]Unlike the Hybrid model of Rudinger et al. (2018), there is no separate classifier parameters for each dataset.

[8]For both datasets, examples from all splits are used, following previous work.

[9]The difference in performance for MEANTIME might come from a difference in splitting: Pouran Ben Veyseh et al.'s (2019) test set has a different size. Some of the gold labels in MEANTIME also seem wrong.

	MegaVeridicality		CB	
	r	MAE	r	MAE
BERT	0.60	1.09	0.59	1.40
Stanovsky et al.	–	–	0.50	2.04
Rudinger et al.	0.64	–	0.33	1.87

Table 5: Performance on MegaVeridicality and CommitmentBank across all splits of the previous model (Stanovsky et al. 2017 and Rudinger et al. 2018) and BERT trained on the concatenation of FactBank, MEANTIME, UW, UDS-IH2 using one set of parameters. White et al. (2018) did not report MAE results for MegaVeridicality.

However, looking at whether BERT improves on the previous state-of-the-art results for its ability to generalize to the linguistic constructions without in-domain supervision, the results are less promising. Table 5 shows performance of BERT trained on four factuality datasets and tested on MegaVeridicality and CB across all splits, and the Rule-based and Hybrid models' performance reported in Jiang and de Marneffe (2019a) and White et al. (2018). BERT improves on the other systems by only a small margin for CB, and obtains no improvement for MegaVeridicality. Despite having a magnitude more parameters and pretraining, BERT does not generalize to the embedded events present in MegaVeridicality and CB. This shows that we are not achieving robust natural language understanding, unlike what the near-human performance on various NLU benchmarks suggests.

Finally, although RoBERTa (Liu et al., 2019) has exhibited improvements over BERT on many different tasks, we found that, in this case, using

pretrained RoBERTa instead of BERT does not yield much improvement. The predictions of the two models are highly correlated, with 0.95 correlation over all datasets' predictions.

6 Quantitative Analysis: Expected Inference

Here, we evaluate our hypothesis that BERT can learn subtle lexical patterns, regardless of whether they align with lexical semantics theories, but struggles when pragmatic reasoning overrides the lexical patterns. To do so, we present results from a quantitative analysis using the notion of expected inference. To facilitate meaningful analysis, we generated two random train/dev/test splits of the same sizes as in Table 3 (besides the standard split) for MegaVeridicality, CB, and RP. All items are present at least once in the test sets. We trained the Multi model using three different random initializations with each split.[10] We use the mean predictions of each item across all initializations and all splits (unless stated otherwise).

6.1 Method

As described above, the expected inference of an item is the factuality label predicted by lexical patterns only. We hypothesize that BERT does well on items where the gold labels match the expected inference, and fails on those that do not.

How to Get the Best Expected Inference? To identify the expected inference, the approach varies by dataset. For the datasets focusing on embedded events (MegaVeridicality, CB, and RP), we take, as expected inference label, the mean labels of training items with the same combination of features as the test item. Theoretically, the signatures should capture the expected inference. However, as shown earlier, the signatures do not always align with the observed annotations, and not all verbs have signatures defined. The mean labels of training items with the same features captures what the common patterns in the data are and what the model is exposed to. In MegaVeridicality and RP, the features are clause-embedding verb, polarity, and frames. In CB, they are verb and entailment-canceling environment.[11]

For FactBank, UW, and MEANTIME, the approach above does not apply because these datasets contain matrix-clause and embedded events. We take the predictions from Stanovsky et al.'s Rule-based model[12] as the expected inference, since the Rule-based model uses lexical rules including the signatures. We omitted UDS-IH2 from this analysis because there are no existing predictions by the Rule-based model on UDS-IH2 available.

Dataset	α	$SE(\alpha)$	β	$SE(\beta)$
FactBank	−0.039	0.018	0.073	0.015
MEANTIME	−0.058	0.033	0.181	0.024
UW	0.004	0.016	0.261	0.016
MegaVeridicality	0.134	0.008	0.142	0.006
CB	0.099	0.020	0.265	0.016
RP	0.059	0.011	0.468	0.012

Table 6: Estimated random intercepts (α) and slopes (β) for each dataset and their standard errors. The fixed intercept is 0.228 with standard error 0.033.

6.2 Results

We fitted a linear mixed effect model using the absolute error between the expected inference and the label to predict the absolute error of the model predictions, with random intercepts and slopes for each dataset. Results are shown in Table 6. We see that the slopes are all positive, suggesting that the error of the expected inference to the label is positively correlated with the error of the model, as we hypothesized.

The slope for FactBank is much smaller than the slopes for the other datasets, meaning that for FactBank, the error of the expected inference does not predict the model's errors as much as in the other datasets. This is due to the fact that the errors in FactBank consist of items for which the lexicalist and crowdsourced annotations may differ. The model, which has been trained on crowdsourced

[10] There is no model performing radically better than the others. The Multi model achieves better results than the Single one on CB and is getting comparable performance to the Single model on the other datasets.

[11] The goal is to take items with the most matching features. If there are no training items with the exact same combination of features, we take items with the next best match, going down the list if the previous features are not available:
- MegaVeridicality and RP: verb-polarity, verb, polarity.
- CB: verb, environment.

[12] https://github.com/gabrielStanovsky/unified-factuality/tree/master/data/predictions_on_test.

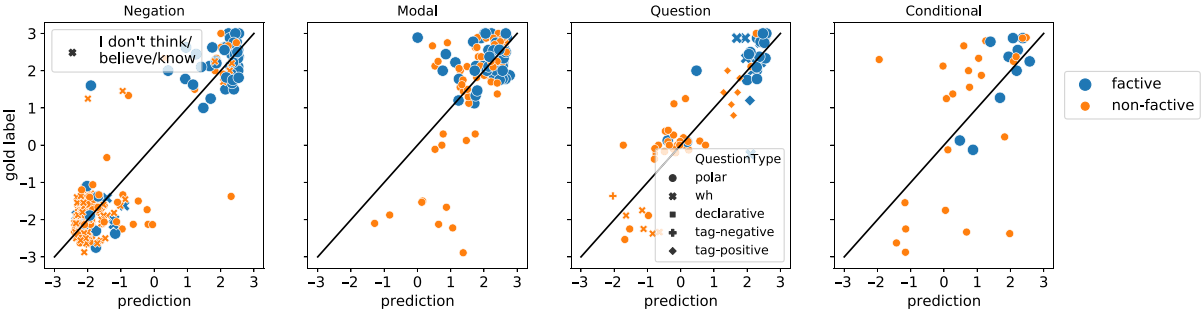

Figure 2: Multi model's predictions compared to gold labels for all CB items present in all splits, by entailment-canceling environment. Diagonal line shows perfect prediction.

datasets, makes predictions that are more in line with the crowdsourced annotations but are errors compared to the lexicalist labels. For example, 44% of the errors are reported events (e.g., *X said that . . .*) annotated as neutral in FactBank (given that both true or false things can be reported) but predicted as factual. Such reported events have been found to be annotated as factual by crowdworkers (de Marneffe et al., 2012; Lee et al., 2015). On the other hand, the expected inference (from the Rule-based model) also follows a lexicalist approach. Therefore labels align well with the expected inference, but the predictions do so poorly.

7 Qualitative Analysis

The quantitative analysis shows that the model predictions are driven by surface-level features. Not surprisingly, when a gold label of an item diverges from the label of items with similar surface patterns, the model does not do well. Here, we unpack which surface features are associated with labels, and examine the edge cases in which surface features diverge from the observed labels. We focus on the CB, RP, and MegaVeridicality datasets because they focus on embedded events well studied in the literature.

7.1 CB

Figure 2 shows the scatterplot of the Multi model's prediction vs. gold labels on CB, divided by each entailment-canceling environment. As pointed out by Jiang and de Marneffe (2019b), the interplay between the entailment-canceling environment and the clause-embedding verb is often the deciding factor for the factuality of the complement in CB. Items with factive embedding verbs tend indeed to be judged as factual (most blue points in Figure 2 are at the top of the panels). ''Neg-raising'' items contain negation in the matrix clause (*not* {*think/believe/know*} ϕ) but are interpreted as negating the content of the complement clause ({*think/believe/know*} *not* ϕ). Almost all items involving a construction indicative of ''Neg-raising'' *I don't think/believe/know* ϕ have nonfactual labels (see $\times$ in first panel of Figure 2). Items in modal environment are judged as factual (second panel where most points are at the top).

In addition to the environment and the verb, there are more fine-grained surface patterns predictive of human annotations. Polar question items with nonfactive verbs often have near-0 factuality labels (third panel, orange circles clustered in the middle). In tag-question items, the label of the embedded event often matches the matrix clause polarity, such as (6) with a matrix clause of positive polarity and a factual embedded event.

(6) [. . .] I **think** it went$^{1.09\ [1.52]}$ to Lockheed, **didn't it**?

Following these statistical regularities, the model obtains good results by correctly predicting the majority cases. However, it is less successful on cases where the surface features do not lead to the usual label, and pragmatic reasoning is required. The model predicts most of the neg-raising items correctly, which make up 58% of the data under negation. But the neg-raising pattern leads the model to predict negative values even when the labels are positive, as in (7).[13]

[13] We use the notation event span$^{\text{label}\ [\text{prediction}]}$ throughout the rest of the paper.

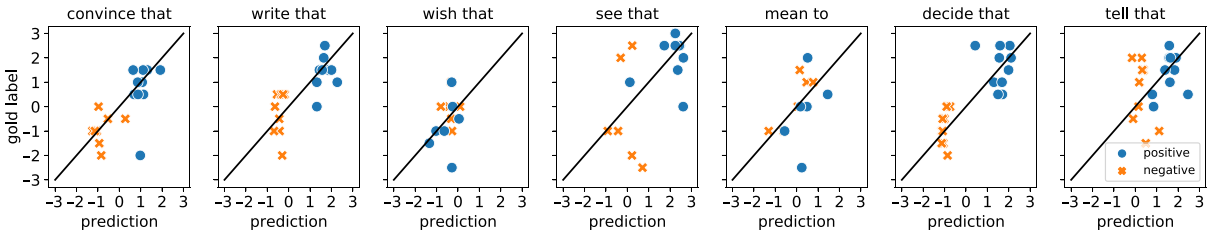

Figure 3: Multi model's predictions compared to gold labels for certain verbs and frames in RP. Diagonal line shows perfect prediction.

(7) [. . .] And I think society for such a long time said, well, you know, you're married, now you need to have your family and I *don't* **think** it's been$^{1.25}$ [-1.99] until recently that they had decided that two people was a family.

It also wrongly predicts negative values for items where the context contains a neg-raising-like substring (*don't think/believe*), even when the targeted event is embedded under another environment: question for (8), antecedent of conditional for (9).

(8) B: All right, well. A: Um, short term, I don't think anything's going to be done about it or probably should be done about it. B: Right. Uh, *are you saying* you *don't* **think** anything should be done in the short term^{0} [-1.73]?

(9) [. . .] I do not believe I am being unduly boastful *if* I **say** that very few ever needed$^{2.3}$ [-1.94] amendment.

7.2 RP

The surface features impacting the annotations in RP are the clause-embedding verb, its syntactic frame, and polarity. Figure 3 shows the scatterplot of label vs. prediction for items with certain verbs and frames, for which we will show concrete examples later. The errors (circled points) in each panel are often far away from the other points of the same polarity on the y-axis, confirming the findings above that the model fails on items that diverge from items with similar surface patterns. Generally, points are more widespread along the y-axis than the x-axis, meaning that the model makes similar predictions for items which share the same features, but it cannot account for variability among such items. Indeed, the mean variance of the predictions for items of each verb, frame, and polarity is 0.19, while the mean variance of the gold labels for these items is 0.64.

Compare (10) and (11): They consist of the same verb *convince* with positive polarity and they have similar predictions, but very different gold labels. Most of the *convince* items of positive polarity are between neutral and factual (between 0 and 1.5), such as (10). The model learned that from the training data: All *convince* items of positive polarity have similar predictions ranging from 0.7 to 1.9, with mean 1.05 (also shown in the first panel of Figure 3). However, (11) has a negative label of -2 unlike the other *convince* items, because the following context *I was mistaken* clearly states that the speaker's belief is false, and therefore the event *they would fetch up at the house in Soho* is not factual. Yet the model fails to take this into account.

(10) I was **convinced** that the alarm was given when Mrs. Cavendish was in the room$^{1.5}$ [1.13].

(11) I was **convinced** that they would fetch up at the house in Soho^{-2} [0.98], but it appears I was mistaken.

7.3 MegaVeridicality

As shown in the expected inference analysis, MegaVeridicality exhibits the same error pattern as CB and RP (failing on items where gold labels differ from the ones of items sharing similar surface features). Unlike CB and RP, MegaVeridicality is designed to rule out the effect of pragmatic reasoning. Thus the errors for MegaVeridicality cannot be due to pragmatics. Where are those stemming from? It is known that some verbs behave very differently in different frames. However, the model was not exposed to the same combination of verb and frame during training and testing, which leads to errors. For example, *mislead*[14] in the *VERBed NP to VP* frame in positive polarity,

[14]Other verbs with the same behavior and similar meaning include *dupe, deceive, fool*.

as in (12), and its passive counterpart (13), suggests that the embedded event is factual (*someone did something*), while in other frame/polarity, the event is nonfactual, as in (14) and (15). The model, following the patterns of *mislead* in other contexts, fails on (12) and (13) because the training set did not contain instances with *mislead* in a factual context.

(12) Someone **misled a particular person to** do$^{2.7}$ [-1.6] a particular thing.

(13) A particular person **was misled to** do$^{2.7}$ [-1.21] a particular thing.

(14) Someone **was misled that** a particular thing happened$^{-1.5}$ [-2.87].

(15) Someone **was*n't* misled to** do$^{-0.3}$ [-0.6] a particular thing.

This shows that the model's ability to reason is still limited to pattern matching: It fails to induce how verb meaning interacts with syntactic frames that are unseen during training. If we augment MegaVeridicality with more items of verbs in these contexts (currently there is one example of each verb under either polarity in most frames) and add them to the training set, BERT would probably learn these behaviors.

Moreover, the model here exhibits a different pattern from White et al. (2018), who found that their model cannot capture inferences whose polarity mismatches the matrix clause polarity, as their model fails on items with verbs that suggest nonfactuality of their complement such as *fake, misinform* under positive polarity. As shown in the expected inference analysis in Section 6, our model is successful at these items, since it has memorized the lexical pattern in the training data.

7.4 Error Categorization

In this section, we study the kinds of reasoning that is needed to draw the correct inference in items that the system does not handle correctly. For the top 10% of the items sorted by absolute error in CB and RP, two linguistically trained annotators annotated which factors lead to the observed factuality inferences, according to factors put forth in the literature, as described below.[15]

[15]This is not an exhaustive list of reasoning types present in the data, and having one of these properties is not sufficient for the model to fail.

Prior Probability of the Event Whether the event described is likely to be true is known to influence human judgments of event factuality (Tonhauser et al., 2018; de Marneffe et al., 2019). Events that are more likely to be factual a priori are often considered as factual even when they are embedded, as in (16). Conversely, events that are unlikely a priori are rated as nonfactual when embedded, as in (17).

(16) [. . .] He took the scuffed leather document case off the seat beside him and banged the door shut with the violence of someone who had *not* **learned** that car doors do not need the same sort of treatment as those of railway carriages$^{2.63}$ [0.96]

(17) In a column lampooning Pat Buchanan, Royko did *not* **write** that Mexico was^{-3} [-0.3] a useless country that should be invaded and turned over to Club Med.

Context Suggests (Non)Factuality The context may directly describe or give indirect cues about the factuality of the content of the complement. In (18), the preceding context *they're French* clearly indicates that the content of the complement is false. The model predicts -0.28 (the mean label for training items with *wish* under positive polarity is -0.5), suggesting that the model fails to take the preceding context into account.

(18) They're French, but **wish** that they were$^{-2.5}$ [-0.28] mostly Caribbean.

The effect of context can be less explicit, but nonetheless there. In (19), the context *which it's mainly just when it gets real, real hot* elaborates on the time of the warnings, carrying the presupposition that the content of the complement *they have warnings here* is true. In (20), the preceding context *Although Tarzan is now nominally in control*, with the marker *although* and *nominally* suggesting that Tarzan is not actually in charge, makes the complement *Kala the Ape-Mom is really in charge* more likely.

(19) [...] B: Oh, gosh, I think I would hate to live in California, the smog there. A: Uh-huh. B: I mean, I *can't* **believe** they have$^{2.33}$ [0.327] warnings here, which it's mainly just when it gets real, real hot.

(20) Although Tarzan is now nominally in control, one does *not* **suspect** that Kala the Ape-Mom, the Empress Dowager of the Jungle, is$^{2.5}$ [-0.23] really in charge.

Discourse Function When sentences are uttered in a discourse, there is a discourse goal or a question under discussion (QUD) that the sentence is trying to address (Roberts, 2012). According to Tonhauser et al. (2018), the contents of embedded complements that do not address the question under discussion are considered as more factual than those that do address the QUD. Even for items that are sentences in isolation, as in RP, readers interpreting these sentences probably reconstruct a discourse and the implicit QUD that the sentences are trying to address. For instance, (21) contains the factive verb *see*, but its complement is labeled as nonfactual (−2).

(21) Jon did *not* **see** that they were^{-2} [1.45] hard pressed.

Such a label is compatible with a QUD asking what is the evidence that Jon has to whether they were hard pressed. The complement does not answer that QUD, but the sentence affirms that Jon lacks visual evidence to conclude that they were hard pressed. In (22), the embedded event is annotated as factual although it is embedded under a report verb (*tell*). However, the sentence in (22) can be understood as providing a partial answer to the QUD *What was the vice president told?*. The content of the complement does not address the QUD, and is therefore perceived as factual.

(22) The Vice President was *not* **told** that the Air Force was trying^{2} [-0.15] to protect the Secretary of State through a combat air patrol over Washington.

Tense/Aspect The tense/aspect of the clause-embedding verb and/or the complement affects the factuality of the content of the complement (Karttunen, 1971b; de Marneffe et al., 2019). In (23), the past perfect *had meant* implies that the complement did not happen (−2.5), whereas in (24) in the present tense, the complement is interpreted as neutral (0.5).

(23) She had **meant** to warn$^{-2.5}$ [-0.24] Mr. Brown about Tuppence.

(24) A bigger contribution **means** to support$^{0.5}$ [1.45] candidate Y.

Subject Authority/Credibility The authority of the subject of the clause-embedding verb also affects factuality judgments (Schlenker 2010, de Marneffe et al., 2012, among others). The subjects of (25), a legal document, and (26), the tenets of a religion, have the authority to require or demand. Therefore what the legal document requires is perceived as factual, and what the tenets do not demand is perceived as nonfactual.

(25) Section 605(b) **requires** that the Chief Counsel gets$^{2.5}$ [0.53] the statement.

(26) The tenets of Jainism do *not* **demand** that everyone must be wearing shoes when they come into a holy place^{-2} [-0.28].

On the other hand, the perceived lack of authority of the subject may suggest that the embedded event is not factual. In (27), although *remember* is a factive verb, the embedded event only receives a mean annotation of 1, probably because the subject *a witness* introduces a specific situational context questioning whether to consider someone's memories as facts.

(27) A witness **remembered** that there were^{1} [2.74] four simultaneous decision making processes going on at once.

Subject-Complement Interaction for Prospective Events Some clause-embedding verbs, such as *decide* and *choose*, introduce so-called ''prospective events'', which could take place in the future (Saurí, 2008). The likelihood that these events will actually take place depends on several factors: the content of the complement itself, the embedding verb, and the subject of the verb. When the subject of the clause-embedding verb is the same as the subject of the complement, the prospective events are often judged as factual, as in (28). In (29), the subjects of the main verb and the complement verb are different, and the complement is judged as neutral.

(28) He **decided** that he must leave no stone unturned$^{2.5}$ [0.43].

(29) Poirot **decided** that Miss Howard must be kept in the dark$^{0.5}$ [1.49].

Even when subjects are the same, the nature of the prospective event itself also affects whether it is perceived as factual. Compare (30) and (31)

both featuring the construction *do not choose to*: (30) is judged as nonfactual whereas (31) is neutral. This could be due to the difference in the extent to which the subject entity has the ability to fulfill the chosen course of action denoted by the embedded predicate. In (30), Hillary Clinton can be perceived to be able to decide where to stay, and therefore when she does not choose to stay somewhere, one infers that she indeed does not stay there. On the other hand, the subject in (31) is not able to fulfill the chosen course of action (where to be buried), since he is presumably dead.

(30) Hillary Clinton does *not* **choose** to stay$^{-2.5}$ [-0.92] at Trump Tower.

(31) He did *not* **choose** to be buried$^{0.5}$ [-0.75] there.

Lexical Inference An error item is categorized under ''lexical inference'' if the gold label is inline with the signature of its embedding verb. Such errors happen on items of a given verb for which the training data do not exhibit a clear pattern because the training items contains items where the verb follows its signature as well as items where pragmatic factors override the signature interpretation. For example, (32) gets a factual interpretation, consistent with the factive signature of *see*.

(32) He did *not* **see** that Manning had glanced^{2} [0.47] at him.

However, the training instances with *see* under negation have labels ranging from -2 to 2 (see the orange $\times$'s in the fourth panel of Figure 3). Some items indeed get a negative label because of the presence of pragmatic factors, such as in (21), but the system is unable to identify these factors. It thus fails to learn to tease apart the factual and nonfactual items, predicting a neutral label that is roughly the mean of the labels of the training items with *see* under negation.

Annotation Error As in all datasets, it seems that some human labels are wrong and the model actually predicts the right label. For instance, (33) should have a more positive label (rather than 0.5), as *realize* is taken to be factive and nothing in the context indicates a nonfactual interpretation.

(33) I did *not* **realize** that John had fought$^{0.5}$ [2.31] with his mother prior to killing her.

	CB		RP	
	#	%	#	%
Prior probability of the event	5	9.1	32	12.8
Context suggests (non)factuality	34	61.8	29	11.6
Question Under Discussion (QUD)			20	8.0
Tense/aspect	1	1.8	8	3.2
Subject authority/credibility	1	1.8	14	5.6
Subject-complement interaction			26	10.4
Lexical inference	12	21.8	88	35.2
Annotation error	2	3.6	33	13.2
Total items categorized	55		250	

Table 7: Numbers (#) and percentages (%) of error items categorized for CB and RP.

In total, 55 items (with absolute errors ranging from 1.10 to 4.35, and a mean of 1.95) were annotated in CB out of 556 items, and 250 in RP (with absolute errors ranging from 1.23 to 4.36, and a mean of 1.70) out of 2,508 items. Table 7 gives the numbers and percentages of errors in each category. The two datasets show different patterns that reflect their own characteristics. CB has rich preceding contexts, and therefore more items exhibit inferences that can be traced to the effect of context. RP has more item categorized under lexical inference, because there is not much context to override the default lexical inference. RP also has more items under annotation errors, due to the limited amount of annotations collected for each item (3 annotations per item).

Although we only systematically annotated CB and RP (given that these datasets focus on embedded events), the errors in the other datasets focusing on main-clause events also exhibit similar inferences as the ones we categorized above, such as effects of context and lexical inference (more broadly construed).[16] Most of the errors concern nominal events. In the following examples—(34) and (35) from UW, and (36) from MEANTIME—model failed to take into account the surrounding context which suggests that the events are nonfactual. In (34), the lexical meaning of *dropped* clearly indicates that the plan is nonfactual. In (35), the death was *faked*, and in (36) production was *brought to an end*, indicating that the death did not happen and there is no production anymore.

(34) In 2011, the AAR consortium attempted to block a drilling joint venture in the Arctic

[16]Some of the error categories only apply to embedded events, including the effect of QUD and subject authority.

between BP and Rosneft through the courts and the plan$^{-2.8}$ [1.84] was eventually dropped.

(35) The day before Raymond Roth was pulled over, his wife, Evana, showed authorities e-mails she had discovered that appeared to detail a plan between him and his son to fake his death$^{-2.8}$ [1.35]

(36) Boeing Commercial Airplanes on Tuesday delivered the final 717 jet built to AirTran Airways in ceremonies in Long Beach, California, bringing production^{-3} [3.02] of McDonnell Douglas jets to an end.

In (37), from FactBank, *just what NATO will do* carries the implication that NATO will do something, and the *do* event is therefore annotated as factual.

(37) Just what NATO will do^{3} [-0.05] with these eager applicants is not clear.

Example (38) from UDS-IH2 features a specific meaning of the embedding verb *say*: Here *say* makes an assumption instead of the usual speech report, and therefore suggests that the embedded event is not factual.

(38) **Say** after I finished$^{-2.25}$ [2.38] those 2 years and I found a job.

Inter-annotator Agreement for Categorization Both annotators annotated all 55 items in CB. For RP, one of the annotators annotated 190 examples, and the other annotated 100 examples, with 40 annotated by both. Among the set of items that were annotated by both annotators, annotators agreed on the error categorization 90% of the time for the CB items and 80% of the time for the RP items. This is comparable to the agreement level in Williams et al. (2020), in which inferences types for the ANLI dataset (Nie et al., 2020) are annotated.

8 Conclusion

In this paper, we showed that, although fine-tuning BERT gives strong performance on several factuality datasets, it only captures statistical regularities in the data and fails to take into account pragmatic factors that play a role on event factuality. This aligns with Chaves's (2020) findings for acceptability of filler-gap dependencies: Neural models give the impression that they capture island constraints well when such phenomena can be predicted by surface statistical regularities, but the models do not actually capture the underlying mechanism involving various semantic and pragmatic factors. Recent work has found that BERT models have some capacity to perform pragmatic inferences: Schuster et al. (2020) for scalar implicatures in naturally occurring data, Jeretič et al. (2020) for scalar implicatures and presuppositions triggered by certain lexical items in constructed data. It is, however, possible that the good performance on those data is solely driven by surface features as well. BERT models still only have limited capabilities to account for the wide range of pragmatic inferences in human language.

Acknowledgment

We thank TACL editor-in-chief Brian Roark and action editor Benjamin Van Durme for the time they committed to the review process, as well as the anonymous reviewers for their insightful feedback. We also thank Micha Elsner, Cory Shain, Michael White, and members of the OSU Clippers discussion group for their suggestions and comments. This material is based upon work supported by the National Science Foundation under grant no. IIS-1845122.

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. https://doi.org/10.1163/9789004253162_004

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

Rui P. Chaves. 2020. What don't RNN language models learn about filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics*, 3(1):20–30.

Gennaro Chierchia and Sally McConnell-Ginet. 1990. *Meaning and Grammar*. MIT Press.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333. https://doi.org/10.1162/COLI_a_00097

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2501

Paloma Jeretič, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.768

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019a. Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1412

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019b. Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6086–6091, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1630

Lauri Karttunen. 1971a. Implicative verbs. *Language*, 47(2):340–358. https://doi.org/10.2307/412084

Lauri Karttunen. 1971b. Some observations on factivity. *Paper in Linguistics*, 4(1):55–69. https://doi.org/10.1080/08351817109370248

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.

Paul Kiparsky and Carol Kiparsky. 1970. Fact. In M. Bierwisch and K. E. Heidolph, editors, *Progress in Linguistics*, pages 143–173. Mouton, The Hague, Paris.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. ROBERTA: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (ICoS-5)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela.

2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.441

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81. Brussels, Belgium, Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1007

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1432

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69. https://doi.org/10.3765/sp.5.6

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1228

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. https://doi.org/10.18653/v1/N18-1067

Roser Saurí. 2008. FactBank 1.0. annotation guidelines.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227. https://doi.org/10.1007/s10579-009-9089-9

Philippe Schlenker. 2010. Local contexts and local meanings. *Philosophical Studies*, 151(1):115–142. https://doi.org/10.1007/s11098-010-9586-0

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the richness of the linguistic signal in predicting pragmatic inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.479

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Proceedings of Semantics and Linguistic Theory 20*. CLC Publications. https://doi.org/10.3765/salt.v20i0.2584

E Allyn Smith and Kathleen Currie Hall. 2014. The relationship between projection and embedding environment. In *Proceedings of the 48th Meeting of the Chicago Linguistics Society*. Citeseer.

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. https://doi.org/10.18653/v1/P17-2056

Judith Tonhauser, David I. Beaver, and Judith Degen. 2018. How projective is projective content? Gradience in projectivity and at-issueness. *Journal of Semantics*, 35(3):495–542. https://doi.org/10.1093/jos/ffy007

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Phil Yeres, Jason Phang, Haokun Liu, Phu Mon Htut, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen,

Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019. `jiant` 1.3: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics. `https://doi.org/10.18653/v1/N18-1101`

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the adversarial natural language inference dataset. *CoRR*, abs/2010.12729.