How well do NLI models capture verb veridicality?

Alexis Ross
Harvard University
alexis_ross@college.harvard.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Abstract

In natural language inference (NLI), contexts are considered *veridical* if they allow us to infer that their underlying propositions make true claims about the real world. We investigate whether a state-of-the-art natural language inference model (BERT) learns to make correct inferences about veridicality in verb-complement constructions. We introduce an NLI dataset for veridicality evaluation consisting of 1,500 sentence pairs, covering 137 unique verbs. We find that both human and model inferences generally follow theoretical patterns, but exhibit a systematic bias towards assuming that verbs are veridical–a bias which is amplified in BERT. We further show that, encouragingly, BERT's inferences are sensitive not only to the presence of individual verb types, but also to the syntactic role of the verb, the form of the complement clause (*to*- vs. *that*-complements), and negation.

1 Introduction

A context is *veridical* when the propositions it contains are taken to be true, even if not explicitly asserted. For example, in the sentence *"He does not know that the answer is 5"*, *"know"* is veridical with respect to *"The answer is 5"*, since a speaker cannot felicitously say the former sentence unless they believe the latter proposition to be true. In contrast, *"think"* would not be veridical here, since *"He does not think that the answer is 5"* is felicitous whether or not it is taken to be true that *"The answer is 5"*. Understanding veridicality requires semantic subtlety and is still an open problem for computational models of natural language inference (NLI) (Rudinger et al., 2018).

This paper deals specifically with veridicality in verb-complement constructions. Prior work in this area has focused on characterizing verb classes–e.g. factives like *"know that"* (Kiparsky and Kiparsky, 1968) and implicatives like *"manage to"* (Karttunen, 1971)–and on incorporating such lexical semantic information into computational models (MacCartney and Manning, 2009). However, increasingly, linguistic evidence suggests that inferences involving veridicality rely heavily on non-lexical information and are better understood as a graded, pragmatic phenomenon (de Marneffe et al., 2012; Tonhauser et al., 2018).

Thus, in this paper, we revisit the question of whether neural models of natural language inference–which are not explicitly endowed with knowledge of verbs' lexical semantic categories–learn to make inferences about veridicality consistent with those made by humans. We solicit human judgements on 1,500 sentence pairs involving 137 verb-complement constructions. Analysis of these annotations provides new evidence of the importance of pragmatic inference in modeling veridicality judgements. We use our collected annotations to analyze how well a state-of-the-art NLI model (BERT, Devlin et al., 2018) is able to mimic human behavior on such inferences. The results suggest that, while not yet solved, BERT represents non-trivial properties of veridicality in context. Our primary contributions are:

- We collect a new NLI evaluation set of 1,500 sentence pairs involving verb-complement constructions (§4).[1]
- We discuss new analysis of human judgements of veridicality and implications for NLI system development going forward (§5).
- We evaluate the state-of-the-art BERT model on these inferences and present evidence that, while there is still work to be done, the

[1] https://github.com/alexisjihyeross/verb_veridicality

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2230–2240,
Hong Kong, China, November 3–7, 2019.

model appears to capture non-trivial properties about verbs' veridicality in context (§6).

2 Background and Related Work

There is significant work, both in linguistics and NLP, on veridicality and closely-related topics (factuality, entailment, etc). We view past work on veridicality within NLP as largely divisible into two groups, which align with two differing perspectives on the role of the NLI task: the sentence-meaning perspective and the speaker-meaning perspective. Briefly, the *sentence meaning* approach to NLI takes the position that NLP systems should strive to model the aspects of a sentence's semantics which are closely derivable from the lexicon and which hold independently of context (Zaenen et al., 2005). In contrast, the *speaker meaning* approach to NLI takes the position that NLP systems should prioritize representation of the goal-directed meaning of a sentence within the context in which it was generated (Manning, 2006). Work on veridicality which aligns with the sentence-meaning perspective tends to focus on characterizing verbs according to their lexical semantic classes (or "signatures"), while work which aligns with the speaker-meaning approach focuses on representing "world knowledge" and evaluating inferences in naturalistic contexts.

Lexical Semantics (Sentence Meaning). Most prior work treats veridicality as a lexical semantic phenomenon. Such work is largely based on lexicons of verb signatures which specify the types of inferences licensed by individual verbs (Karttunen, 2012; Nairn et al., 2006; Falk and Martin, 2017). White and Rawlins (2018); White et al. (2018) evaluated neural models' ability to carry out inferences in line with these signatures, making use of templatized "semantically bleached" stimuli (e.g. *"someone knew something"*) in order to avoid confounds introduced by world knowledge and pragmatic inference. McCoy et al. (2019) perform a similar study, though without specific focus on veridicality lexicons.

Most applied work related to veridicality also falls under the lexical semantic approach. In nearly all cases, relevant system development involves explicit incorporation of verb lexicons and associated logical inference rules. MacCartney and Manning (2009); Angeli and Manning (2014); and others incorporated knowledge of verb signatures within a natural logic framework (MacCartney, 2009; Sánchez Valencia, 1991) in order to perform natural language inference. Richardson and Kuhn (2012) incorporated signatures into a semantic parsing system. Several recent models of event factuality similarly make use of veridicality lexicons as input to larger machine-learned systems for event factuality (Saurí and Pustejovsky, 2012; Lotan et al., 2013; Stanovsky et al., 2017; Rudinger et al., 2018). Cases et al. (2019) used nested veridicality inferences as a test case for a meta-learning model, again assuming verb signatures as "meta information" known *a priori*.

Pragmatics (Speaker Meaning). Geis and Zwicky (1971) observed that implicative verbs often give rise to "invited inferences", beyond what is explainable by the lexical semantic type of the verb. For example, on hearing *"He did not refuse to speak"*, one naturally concludes that *"He spoke"* unless additional qualifications are made (e.g. *"...he just didn't have anything to say"*). de Marneffe et al. (2012) explored this idea in depth and presented evidence that such pragmatic inferences are both pervasive and annotator-dependent, but nonetheless systematic enough to be relevant for NLP models. Karttunen et al. (2014) makes similar observations specifically in the case of evaluative adjectives, and Pavlick and Callison-Burch (2016) specifically in the case of simple implicative verbs. In non-computational linguistics, Simons et al. (2017, 2010); Tonhauser et al. (2018) take a strong stance and argue that veridicality judgements are entirely pragmatic, dependent solely on the question under discussion (QUD) within the given discourse.

This Work. This paper assumes the speaker-meaning approach: we take the position that models which consistently mirror human inferences about veridicality in context can be said to understand veridicality in general. We acknowledge that the question of what is the "right" approach to NLI has existed since the original definition of the recognizing textual entailment (RTE) task (Dagan et al., 2006) and remains open. However, there has been a *de facto* endorsement of the speaker-meaning definition, evidenced by the widespread adoption of NLI datasets which favor informal, "natural" inferences over prescriptivist annotation guidelines (Manning, 2006; Bowman et al., 2015; Williams et al., 2018). (Note, recently, there have been explicit endorsements as well; see Westera

Factive [+/+]	He **realized that** he had to leave this house. He **did not realize that** he had to leave this house.	→ →	He had to leave this house. He had to leave this house.
Implic. [+/−]	At that moment, I **happened to** look up. At that moment, I **did not happen to** look up.	→ → ¬	At that moment, I looked up. At that moment, I looked up.
Implic. [−/○]	He **refused to** do the same. He **did not refuse to** do the same.	→ ¬ ↛	He did the same. He did the same.
NA [○/○]	Many **felt that** its inclusion was a mistake. Many **did not feel that** its inclusion was a mistake.	↛ ↛	Its inclusion was a mistake. Its inclusion was a mistake.

Table 1: Examples of several verb signatures and illustrative contexts for each. Signature $s1/s2$ denotes that the complement will project with polarity $s1$ in a positive environment and polarity $s2$ in a negative environment.

and Boleda (2019)). Thus, from this perspective, we ask: do NLI models which are *not* specifically endowed with lexical semantic knowledge pertaining to veridicality nonetheless learn to model this semantic phenomenon?

3 Projectivity and Verb Signatures

Veridicality is typically treated as a lexical semantic property of verbs, specified by the verb's *signature*. These signatures can indicate that a verb licenses positive (+), negative (−), or neutral (○) inferences. Specifically, Karttunen (2012) defines these as two-bit signatures, to reflect that verbs[2] may behave differently in positive vs. negative environments. For example, a factive verb construction like *"know that"* has a +/+ signature, indicating that the complement projects positively in both positive and negative environments. That is, both *"He knows that the answer is 5"* and *"He does not know that the answer is 5"* imply that *"The answer is 5"*. In contrast, a verb like *"manage to"* has the signature +/− since, in a positive environment, the complement projects (*"I managed to pass"*→*"I passed"*) but, in a negative environment, the *negation* of the complement projects (*"I did not manage to pass"*→ ¬*"I passed"*). Other verbs may exhibit veridicality only in positive or negative environments but not in both. For example, *"refuse to"* has signature −/○: *"She refused to dance"*→ ¬*"She danced"*, but *"She did not refuse to dance"* neither implies nor contradicts the claim *"She danced"*. Still other verbs are entirely non-veridical (○/○). For example, *"hope to"* is not expected to license any inferences about the truth of its complement. We consider 8 signatures[3] in total. Table 1 provides several examples. Table 2 lists all of the signatures and the corresponding verbs we consider.

4 Data

For our analysis, we collect an NLI dataset for veridicality evaluation derived from the MNLI corpus. This data is publicly available at `https://github.com/alexisjihyeross/verb_veridicality`.

4.1 Generating NLI Pairs

We generate NLI-style premise/hypothesis pairs based on sentences drawn from the train+dev splits of the MultiNLI (Williams et al., 2018) corpus. Specifically, we collect all sentences appearing in MNLI[4] which contain any verb-complement construction, e.g. any sequence matching the pattern: `verb` {*"to"*|*"that"*} {VP|S}. Since we aim to manipulate the environment (positive vs. negative) ourselves in a controlled manner, we filter out sentences which already contain explicit negation words (e.g. *"no"*, *"not"*, *"never"*, *"n't"*), conditionals (*"if"*), and passive constructions (*"was intended to"*). From the selected sentences, we take a stratified sample over the lemma of `verb`. For use in later analysis, we associate each lemma with a signature using a manually-curated dictionary[5] of implicative and factive verbs, and assign the signature of ○/○ to verbs which do not appear in the dictionary. We sample a set of sentences for every lemma, taking

[2] Karttunen (2012) actually discusses these signatures for implicative verbs only. We adopt the notation, but use it in for factives and uncategorized verbs as well.

[3] Seven from `http://web.stanford.edu/group/csli_lnr/Lexical_Resources/`, plus ○/○.

[4] We removed sentences from the telephone genre, since the parses on these tended to be noisy and thus our manipulations often yielded incorrect or difficult to interpret sentences.

[5] `http://web.stanford.edu/group/csli_lnr/Lexical_Resources/`

no more than 40 sentences per lemma, and weighting our sampling to prefer shorter sentences, in order to reduce cognitive load on our raters.

$+/+$ realize that (34) know that (32) remember that (17) find that (12) notice that (12) reveal that (12) acknowledge that (11) admit that (11) learn that (11) observe that (11) see that (11) note that (10) recognize that (10) understand that (10) discover that (8) $+/-$ manage to (30) begin to (12) serve to (11) start to (11) dare to (8) use to (7) get to (6) come to (5) $-/+$ forget to (15) fail to (10) $\circ/+$ suspect that (11) explain that (10) mean to (10) predict that (10) $\circ/-$ attempt to (28) $-/\circ$ refuse to (36) decline to (12) remain to (7) $+/\circ$ show that (12) confirm that (11) demonstrate that (10) ensure that (9) help to (9) tend to (8) $\circ/\circ$ try to (34) hope that (20) hope to (18) mention that (14) like to (12) continue to (12) expect that (12) agree that (12) love to (12) reply that (12) conclude that (12) say that (12) complain that (12) speculate that (12) state that (12) suggest that (12) worry that (12) mean that (12) intend to (11) insist that (11) imply that (11) indicate that (11) plan to (11) promise to (11) prove to (11) saw that (11) seem that (11) tell that (11) think that (11) felt that (11) write that (11) decide to (11) assume that (11) believe that (11) assert that (11) concern that (11) estimate that (11) convince that (11) decide that (11) appear that (11) argue that (11) aim to (11) cease to (10) strive to (10) proceed to (10) choose to (10) seem to (10) prove that (10) provide that (10) seek to (10) appear to (10) comment that (10) contend that (10) want to (10) doubt that (10) feel that (10) fear that (10) agree to (10) announce that (9) claim that (9) struggle to (9) hear that (9) propose to (9) wish to (9) say to (9) turn to (8) wish that (8) work to (8) advise that (8) move to (8) claim to (8) expect to (8) report that (8) happen to (8) propose that (8) hold that (8) declare that (8) prefer to (8) need to (8) give that (7) deserve to (7) threaten to (7) exist to (7) be that (7) prepare to (6) wait to (6) pretend to (6) ask to (6) return to (6) request that (5) demand that (4) recommend that (4) require that (4)

Table 2: 137 verbs belonging to 8 signatures. Parentheses denote number of contexts in which each verb appears in our final, annotated dataset (§4)

.

We consider each sampled sentence S to be a candidate premise. We then generate premise/hypothesis pairs as follows. We use the parse tree provided by MNLI to extract the complement clause C. When needed, we inflect[6] the verb in the complement to match the tense of the main verb. We then generate two $\langle p, h \rangle$ pairs: the sentence and the complement as-is $\langle S, C \rangle$, and the negated sentence plus the complement $\langle \neg S, C \rangle$. For example, given an original sentence like *"He knows that the answer is 5"*, we would generate two $\langle p, h \rangle$ pairs: $\langle$*"He knows that the answer is 5"*, *"The answer is 5"*$\rangle$ and $\langle$*"He does not know that the answer is 5"*, *"The answer is 5"*$\rangle$. The examples shown in Table 1 illustrate $\langle p, h \rangle$ pairs drawn from our dataset, generated this way.

4.2 Annotation

For each $\langle p, h \rangle$ pair, we collect human judgements on Amazon Mechanical Turk. We have raters label entailment on a 5-point likert scale in which -2 means that h is definitely *not* true given p and 2 means that h is definitely true given p. This ordinal labelling scheme[7] matches prior work on common sense inference (Zhang et al., 2017), and on veridicality specifically (de Marneffe et al., 2012). We do not provide examples for boundary cases (the difference between $-2, -1$ or $1, 2$) to avoid biasing raters by providing explicit guidance about the extent to which common sense can factor in. Raters have the option of indicating with a check box that one or both sentences does not make sense, and thus that they are unable to judge. We require that raters have had at least 100 approved tasks, have maintained an 98% approval rating, and are located in an primarily English-speaking country (US, AU, GB, CA). We collect three annotations per p/h pair, and pay \$0.10 per set of six pairs labelled.

Quality Controls and Exclusion Criteria. We remove all sentence pairs in which one or more raters checked the "does not make sense" box. We remove sentences from our analysis unless both the $\langle S, C \rangle$ and the $\langle \neg S, C \rangle$ pairs passed this filter. Finally, we remove verbs from our analysis which, after the above filtering, do not appear with at least 4 sentences (i.e. 8 p/h pairs). Our final dataset contains 137 verb types across 1,498 sentences (2,996 pairs). Table 2 lists the verbs included in our dataset and the number of sentences in which each appears.[8] To measure inter-rater agreement, for each example and each of the three raters assigned to the example, we calculated the correlation between that rater's score and the averaged score of the other two raters. The Spearman correlation among raters, averaged across the three raters for each example, was 0.78 for positive contexts and 0.74 for negative contexts.

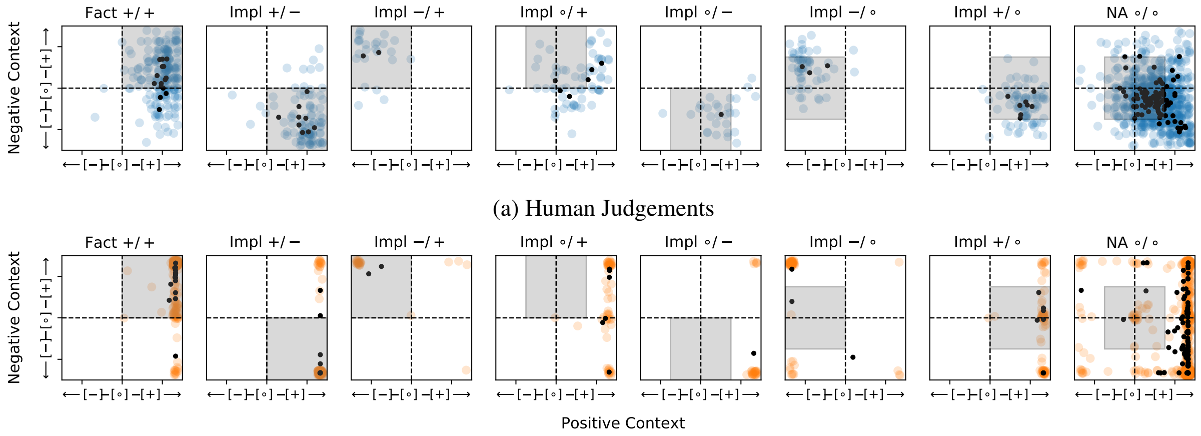
4.3 Aggregation

We take the mean of the three human judgements for each sentence pair. We then represent each verb v (in the context of a given sentence S) using a continuous analog of the projectivity signatures discussed in §3. That is, we take the mean score for $\langle S, C \rangle$ as a measure of the veridicality of v (in the context of S) in a positive environment, and the mean score for $\langle \neg S, C \rangle$ as a measure of the veridicality in a negative environment. For example, given S=*"David Plotz failed to change my mind"* and C=*"David Plotz changed my mind"*, we get a soft projectivity "signature" of $-2.0/1.67$, which is consistent with the expected (discrete) $-/+$ signature for *"fail to"*.

[6] https://www.clips.uantwerpen.be/pages/pattern-en

[7] Full annotation guidelines in Supplementary.

[8] See Supplementary for number of contexts that were excluded for each verb.



(a) Human Judgements

(b) Model Predictions

Figure 1: Human judgements (top, blue) and model predictions (bottom, orange) for verbs in each category. Gray squares denote the region in which judgements are expected fall, given the signature. Each colored dot corresponds to a single context (verb within a specific sentence); each black dot corresponds to a single verb (averaged score over all contexts in which it was judged).

5 Analysis of Human Judgements

Figure 1a plots these soft veridicality signatures for each sentence. We see that, averaged across all the contexts in which they are judged, verbs tend to behave as expected given their assigned lexical semantic signature. However, we observe two noteworthy trends, discussed below. We note that these observations are consistent with arguments made by de Marneffe et al. (2012) about the strong effects of pragmatics on veridicality judgments.

Veridicality Bias. First, we observe a systematic "veridicality bias", in which inferences about complements are often made (positive or negative), even in environments when the expectation is that the verb is non-veridical (∘ signature). This trend is most evident in the case of verbs with ∘/∘ signatures, for example, *"think that", "want to"*. While embedding under such verbs should not license any inferences about the truth of the complement, we observe that, in practice, these verbs tend to behave like +/− verbs. That is, the complement is taken as true in positive environments and as false in negative environments. Table 3 shows some examples for which this is the case.

Within-Verb Variation. Second, we observe that, while individual verb types tend to behave in line with their expected signatures *on average*, signatures provide a weak signal for predicting the inferences licensed by the verb in any sentence individually. That is, within each signature, we see high variance across contexts, in all cases spanning at least 2 points (on our −2 to 2 scale). Table 4 shows examples of words receiving different signatures based on context. Quantitatively, in an ordinary least squares regression[9], we find that using verb signature alone to predict the human judgments in a given context explains only a small amount of the observed variation ($R^2 \approx 0.11$).[10]

[+] (1.7)	The GAO has **indicated that** it is unwilling to compromise. → It is unwilling to compromise.
[−] (−1.0)	The GAO has **not indicated that** it is unwilling to compromise. → ¬ It is unwilling to compromise.
[+] (1.3)	But most visitors **prefer to** linger in Formentera. → But most visitors linger in Formentera.
[−] (−1.3)	But most visitors do not **prefer to** linger in Formentera. → ¬ But most visitors linger in Formentera.

Table 3: Examples of verbs which are expected to be ∘/∘, but which behave like +/− in context. We refer to this trend as a general "veridicality bias".

[9] `statsmodels.regression.linear_model.OLS.html`

[10] For context, using the verb type itself produced $R^2 \approx 0.72$. We experimented with other contextual features in combination with linguistic category and/or verb type (e.g. tense of the main verbs, first vs. third person subjects, etc.) to try to improve the fit of the model, but did not find any note-

[+] (1.7)	Everyone **knows that** the CPI is the most accurate. → The CPI is the most accurate.
[+] (1.7)	Everyone **does not know that** the CPI is the most accurate. → The CPI is the most accurate.
[+] (0.7)	I **know that** I was born to succeed. → I was born to succeed.
[∘] (0.3)	I **do not know that** I was born to succeed. ↛ I was born to succeed.

Table 4: Examples of how the factive verb *"know that"* can exhibit different signatures, depending on context.

Takeaways. Overall, we interpret the above analysis as evidence that veridicality judgments rely heavily on contextual as opposed to purely lexical semantic factors. While this is not a novel conclusion (Simons et al., 2010; de Marneffe et al., 2012), it is still frequently the case that system development concerned with improving veridicality judgements nearly always proceeds by incorporating explicit lexical semantic knowledge into the pipeline or architecture (Richardson and Kuhn, 2012; Lotan et al., 2013; Stanovsky et al., 2017; MacCartney and Manning, 2009; Angeli and Manning, 2014; Saurí and Pustejovsky, 2012; Cases et al., 2019; Rudinger et al., 2018). Our analysis suggests such approaches are likely to yield only incremental gains. While admittedly more difficult to encode, focusing on context-specific factors first, e.g. predicate classes and pragmatics (de Marneffe et al., 2012) or question under discussion (Simons et al., 2010), would likely be more productive and may ultimately override the need for verb signatures altogether.

6 Analysis of BERT Predictions

We now turn to our primary question: do current NLI models capture the veridicality of verbs? In particular, we are interested in the behavior of a distributional model that is not specifically endowed with lexical semantic information related to veridicality. We ask two questions. First: does such a model learn to make inferences consistent with those made by humans? Second: if the model does mirror human inferences, are the predictions based solely on the presence of specific lexical items, or are they sensitive to structural factors (namely, syntactic position and complement type)? Again, we prioritize modeling speaker meaning. Thus, we believe the model should ideally reflect the same biases and variation observed in the human judgments, not necessarily the inferences expected based on the lexical semantic signatures of the verbs.

6.1 Setup

We use the state-of-the-art BERT-based NLI model. Specifically, we use the original TensorFlow implementation[11] of the NLI model built on top of the pertained BERT language model (Devlin et al., 2018). We use the model off-the-shelf, with default training setup and hyperparameters. To fine-tune the model for the NLI task, we use the standard train/dev splits from the MNLI corpus, but, to avoid confounds, we remove the 1,500 p/h pairs from which our new test set is derived (as described in §4).[12] The model is trained to make a softmax classification over three classes: {ENTAILMENT, CONTRADICTION, NEUTRAL}. When necessary to compare these discrete predictions to our continuous human judgments, we map the prediction to a continuous value using $P(\text{ENTAILMENT}) - P(\text{CONTRADICTION})$. This score ranges from -1 to 1 and is comparable to our human scores. Conversely, when necessary to compare the continuous human score to the discrete predictions, we discretize scores into evenly-sized bins[13]. Overall, similar performance trends hold whether we compare in discrete space or continuous space. In the analyses below, we use whichever is most interpretable, as reported.

6.2 Overall Prediction Performance

We first measure raw prediction performance: do the inferences made by the BERT mirror the inference that our human raters made? Figure 1b shows scatter plots of the models predictions (mapped into continuous space) side-by-side with the hu-

worthy effects. It is likely that more careful featurization of highly relevant concepts–e.g. at-issueness (Tonhauser et al., 2018)–could yield more conclusive insights about which aspects of context lead to within-signature, or within-verb, variation. We leave such analysis for future work, and conclude simply that endowing NLI models with knowledge of projectivity signature is not alone sufficient for producing human-like inferences on such sentences.

[11] https://github.com/google-research/bert

[12] See Supplementary Material for a discussion of the prevalence of veridicality phenomena in MNLI train.

[13] CON: $s < -\frac{2}{3}$; NEU: $-\frac{2}{3} \leq s < \frac{2}{3}$; ENT: $\frac{2}{3} \leq s$

	Count		Positive			Negative			
	Sent.	Verb	Exp.	Acc.	ρ	Exp.	Acc.	ρ	Example Verbs
Fact	212	15	+	0.62	0.17	+	0.29	0.40	realize that, know that
Impl	100	9	+	0.57	0.51	−	0.73	0.51	manage to, begin to
Impl	25	2	−	0.80	0.61	+	0.52	0.39	forget to, fail to
Impl	63	6	○	0.27	0.21	+	0.43	0.43	suspect that, explain that
Impl	28	1	○	0.11	0.25	−	0.71	0.45	attempt to
Impl	55	3	−	0.93	0.70	○	0.02	-0.10	refuse to, decline to
Impl	80	8	+	0.38	0.21	○	0.54	0.21	show that, confirm that
NA	935	93	○	0.21	0.35	○	0.44	0.47	try to, hope to
Overall	1,498	137		0.34	0.63		0.44	0.57	

Table 5: Accuracy and Spearman correlation of BERT MNLI model predictions against human judgements. The +/ − /○ symbols denote the expected labels based on the lexical semantic category of the verb, and are not necessarily the labels given by our human annotators (compare against Figure 1).

man scores just discussed. Table 5 shows performance evaluated against human judgements in terms of both (discrete) classification accuracy and (continuous) Spearman correlation. Broadly speaking, the model's predictions appear to follow the same trends as the humans' ratings (Figure 1). That is, averaged across contexts, the model's treatment of verbs is the same as the humans' treatment: largely in line with the signatures, but with a bias against assuming neutral (non-veridical) behavior. However, whereas the humans' judgments span all levels of certainty (taking a range of values from -2 to 2), the model tends to make predictions with high confidence. This is especially the case in positive environment, where the model nearly always predicts with $99+\%$ confidence. In negative environments, the model expresses a grater range of uncertainty values, and is much more closely in line with what we observe in human judgements. In terms of quantitative measures of accuracy (Table 5), the most notable trend is that the model performance is highest for cases in which the *negation* of the complement is expected to project (− signatures). This is true regardless of whether that behavior occurs in a positive or negative environment. We note that, for such cases, human judgements closely align with the lexical semantic predictions. The model performs worst in positive environments when the verb is expected to be non-veridical (○ signatures). This appears to result from the model's tendency to over-exaggerate the veridicality bias: i.e. whereas humans show a general tendency to assume the complement projects in these cases, the model predicts ENTAILMENT with near certainty (see Figure 1).

6.3 Counterfactual Analysis

Next, we ask: are the above-observed trends in BERT's predictions driven predominantly by lexical priors–i.e. the presence of a specific verb–or are they sensitive to other lexicosyntactic factors that should ideally affect the inference?

Experimental Design. For each verb construction *vt* (e.g. *"try to"* or *"realize that"*) in our dataset, we perform several manipulations in which we insert *v* or *t* into sentences where they did not originally appear, and observe the effect this has on the distribution of the model's predictions. Our specific manipulations and the expected effects are described below. Table 6 shows examples. For convenience, we use D to refer to the set of all the $\langle S, C\rangle$ pairs in our dataset, D_{vt} to refer to all the pairs in which *vt* appears as the main verb clause in S, and D_{to} (D_{that}) to refer to all the pairs in which C is a *to*-complement (*that*-complement). When clear from context, we abuse notation and use e.g. D to refer both to the dataset itself and to the distribution of the model's predictions when run over the dataset.

Replace Main Verb: For each pair $\langle S, C\rangle \in D$, we replace the main verb in S with the target verb *v*, generating a new premise S^*. We expect that, if the model is sensitive specifically to presence of *v* and its effect on inferences, then the distribution of model predictions over all $\langle S^*, C\rangle$ pairs should be more similar to the target distribution of predictions over all of D_{vt} than to the baseline distribu-

	Main Verb (Match)		Main Verb (Mismatch)
S	He **attempted to** overcome the sensation.	S	I **decided that** the department had acted illegally.
S*	He **tried to** overcome the sensation.	S*	I **tried that** the department had acted illegally.
C	He overcame the sensation.	C	The department had acted illegally.
	Complement Verb		Complement Type
S	He attempted to **overcome** the sensation.	S	They **tried to** get his attention.
S*	He attempted to **try** the sensation.	S*	They **tried that** get his attention.
C*	He **tried** the sensation.	C	They got his attention.

Table 6: Examples of counterfactual manipulations with the target verb construction *“try to”*.

tion of predictions over all of D. We differentiate between settings with “matched” complement types, where we generate S^* from pairs in D_t, from those with “mismatched” complement types, where we generate from pairs in $D_{\neg t}$. E.g., for a target vt = *“try to”*, we consider substitutions into premises from D_{to} as “matched” and substitutions into D_{that} as “mismatched”. Preserving this distinction allows us to both avoid confounds due to ungrammatical substitutions, and to investigate whether the model is sensitive to verbs which behave differently when they take different complements. For example, *“forget”* is $-/+$ when it takes *to* but $+/+$ when it takes *that*.

Replace Complement Verb: For each $\langle S, C \rangle \in D$, we replace the main verb in C with the target verb *v*, generating a new hypothesis C^*. We expect that, if the model is sensitive not just to the presence of *v*, but also its syntactic role, then the distribution of predictions over all $\langle S, C^* \rangle$ pairs should resemble the baseline distribution over D more than the target distribution over D_{vt}.

Replace Complement Type: For each pair $\langle S, C \rangle \in D_{vt}$, we replace the t in S with the alternative complement type (*“to”* → *“that”*; *‘that”* → *“to”*), generating a new premise S^*. This generates ungrammatical sentences, and serves as a control experiment, to check whether the model is considering the entirety of the context in which the verb construction appears, or merely the vt bigram. We expect that, if the model is considering the whole context, the distribution of predictions over all $\langle S^*, C \rangle$ should resemble the target distribution D_{vt} more than the baseline distribution D_t.

Results. Table 7 shows, for verbs within each signature, the KL divergence between the post-manipulation prediction distribution (D^*) and 1) the baseline distribution (D) and 2) the target distribution (D_{vt}). Results are shown for both the main and complement verb manipulations.

A few trends are worth highlighting. First, we do see evidence that the model's prediction depends at least in part on the individual verb type. This is supported by the fact that, across verb signatures, manipulation of the main verb leads to distributions which are more similar to the target verb distribution D_{vt} than to the baseline distribution D. This trend is strongest for verbs which involve $-$ signatures. Second, we see encouraging, though not overwhelming, evidence that the model's prediction are sensitive to the syntactic position of the verb. This is supported by the fact that, in general, the similarity between D^* and D_{vt} is much lower (higher KL) when the manipulation occurs in the complement clause compared than when it occurs in the main clause. Note that, ideally, this manipulation should not effect the prediction distribution at all. Nonetheless, the trend is clear and points in the right direction.

Table 8 shows the KL divergence between D^* and the target verb distribution D_{vt} in the matched and mismatched cases.[14] We see that BERT behaves as we hope–namely, it makes different predictions for *v to* constructions and *v that* constructions, even when the *v* is the same. Manipulating the main verb only substantially affects predictions when the manipulation occurs in a context with the right complement type (matched); when the manipulation results in a ungrammatical sentence (mismatched), the prediction remains close to baseline. An example of such verb-construction differentiation is shown in Figure 2 for the verb *know that*, but this is a trend seen across verbs. Moreover, we see that this effect is not just driven by sensitivity to the specific *vt* bigram. That is, simply swapping *“to”* with *“that”* (or vice-versa) in a naturally-occurring context leads to a small

[14] See Supplementary for breakdown by verb type.

		Main Verb		Compl. Verb	
		Pos.	Neg.	Pos.	Neg.
$+/+$	$D^*\|D_{vt}$	0.00	0.01	0.02	0.14
	$D^*\|D$	0.17	0.43	0.05	0.08
$+/-$	$D^*\|D_{vt}$	0.01	0.02	0.17	0.01
	$D^*\|D$	0.17	0.67	0.04	0.39
$-/+$	$D^*\|D_{vt}$	0.04	0.06	1.36	1.39
	$D^*\|D$	1.32	1.40	0.04	0.42
$\circ/+$	$D^*\|D_{vt}$	0.01	0.02	0.01	0.06
	$D^*\|D$	0.12	0.22	0.03	0.05
$\circ/-$	$D^*\|D_{vt}$	0.05	0.07	0.13	0.01
	$D^*\|D$	0.01	0.76	0.04	0.42
$-/\circ$	$D^*\|D_{vt}$	0.28	0.11	1.90	0.53
	$D^*\|D$	0.65	0.26	0.03	0.35
$+/\circ$	$D^*\|D_{vt}$	0.00	0.06	0.05	0.31
	$D^*\|D$	0.14	0.04	0.00	0.02
$\circ/\circ$	$D^*\|D_{vt}$	0.00	0.02	0.01	0.06
	$D^*\|D$	0.02	0.00	0.00	0.02

Table 7: Comparison (KL divergence) of post-manipulation prediction distribution to target verb distribution (D_{vt}, top row) and baseline distribution (D, bottom row). High similarity to D_{vt} suggests the model changed its predictions in response to the manipulation.

shift in the distribution of the model's predictions away from the target D_{vt} distribution, but not to the same degree as replacing a *"that"*-taking verb with a *"to"*-taking verb in a naturally-occurring *"that"* context (or vice-versa). This result provides some evidence that BERT's prediction is influenced by aspects of the context other than just the presence of the vt bigram.

	Match	Mis.	Swap
Fact $+/+$	0.01	0.48	0.25
Impl $+/-$	0.01	0.86	0.16
Impl $-/+$	0.05	1.72	0.10
Impl $\circ/+$	0.01	0.16	0.22
Impl $\circ/-$	0.06	0.86	0.00
Impl $-/\circ$	0.19	1.16	0.71
Impl $+/\circ$	0.03	0.23	0.28
NA $\circ/\circ$	0.01	0.03	0.46

Table 8: KL divergence between D^* and D_{vt} for complement type manipulations (*"to"* vs. *"that"*). Inserting v into a context affects BERT's predictions only when the complement is compatible with v.

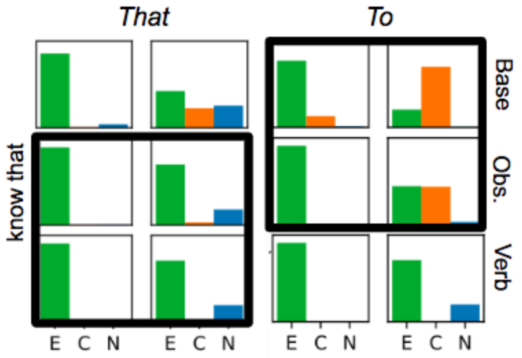


Figure 2: Results for the main verb replacement experiment for the verb *know that*. In examples with the *VB that* construction, the replacement results in predictions that resemble those for *know that*. For *VB to* examples, the resulting counterfactual predictions resemble those of the base examples.

7 Conclusion

We investigate how well BERT, a neural NLI model not explicitly endowed with knowledge of lexical semantic verb signatures, is able to learn to make correct inferences about veridicality. We collect a new NLI dataset of human veridicality judgements. We observe that human judgments often differ from what is predicted given the lexical semantic types of verbs, and that BERT is able to replicate many of these judgments, although there is still significant room for improvement. Through counterfactual experiments, we show that individual verbs strongly influence BERT's predictions, and that these cues interact with syntactic information in desirable ways.

Acknowledgments

Thank you to Sam Bowman and Ben Van Durme for helpful feedback early on in this project. Thank you to the rest of the JSALT 2018 Sentence Representation Team for support throughout. Much of this work was conducted at the 2018 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, and supported by Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, Microsoft and Mitsubishi Electric Research Laboratories.

References

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 534–545.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ignacio Cases, Clemens Rosenbaum, Matthew Riemer, Atticus Geiger, Tim Klinger, Alex Tamkin, Olivia Li, Sandhini Agarwal, Joshua D Greene, Dan Jurafsky, et al. 2019. Recursive routing networks: Learning to compose modules for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ingrid Falk and Fabienne Martin. 2017. Towards an inferential lexicon of event selecting predicates for French. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Michael L Geis and Arnold M Zwicky. 1971. On invited inferences. *Linguistic inquiry*, 2(4):561–566.

Lauri Karttunen. 1971. Implicative verbs. *Language*, pages 340–358.

Lauri Karttunen. 2012. Simple and phrasal implicatives. In **SEM 2012: 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 (SemEval 2012)*, pages 124–131, Montréal, Canada. Association for Computational Linguistics.

Lauri Karttunen, Stanley Peters, Annie Zaenen, and Cleo Condoravdi. 2014. The chameleon-like nature of evaluative adjectives. *Empirical Issues in Syntax and Semantics*, 10:233–250.

Paul Kiparsky and Carol Kiparsky. 1968. *Fact*. Linguistics Club, Indiana University.

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.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Citeseer.

Bill MacCartney and Christopher D Manning. 2009. *Natural language inference*. Citeseer.

Christopher D Manning. 2006. Local textual inference: its hard to circumscribe, but you know it when you see it–and nlp needs it.

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.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

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)*.

Ellie Pavlick and Chris Callison-Burch. 2016. Tense manages to predict implicative behavior in verbs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2225–2229.

Kyle Richardson and Jonas Kuhn. 2012. Light textual inference for semantic parsing. In *Proceedings of COLING 2012: Posters*, pages 1007–1018, Mumbai, India. The COLING 2012 Organizing Committee.

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: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Víctor Manuel Sánchez Valencia. 1991. *Studies on natural logic and categorial grammar*. VM Sanchez Valencia.

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, David Beaver, Craige Roberts, and Judith Tonhauser. 2017. The best question: Explaining the projection behavior of factives. *Discourse Processes*, 54(3):187–206.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Semantics and linguistic theory*, volume 20, pages 309–327.

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.

Judith Tonhauser, David I Beaver, and Judith Degen. 2018. How projective is projective content? gradience in projectivity and at-issueness.

Matthijs Westera and Gemma Boleda. 2019. Dont blame distributional semantics it cant do entailment. In *Proceedings of the 13th International Conference on Computational Semantics*.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society, to appear. Amherst, MA: GLSA Publications*.

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.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 31–36. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.