# Evaluating BERT for natural language inference: A case study on the CommitmentBank

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

## Abstract

Natural language inference (NLI) datasets (e.g., MultiNLI) were collected by soliciting hypotheses for a given premise from annotators. Such data collection led to annotation artifacts: systems can identify the premise-hypothesis relationship without observing the premise (e.g., negation in hypothesis being indicative of contradiction). We address this problem by recasting the CommitmentBank for NLI, which contains items involving reasoning over the extent to which a speaker is committed to complements of clause-embedding verbs under entailment-canceling environments (conditional, negation, modal and question). Instead of being constructed to stand in certain relationships with the premise, hypotheses in the recast CommitmentBank are the complements of the clause-embedding verb in each premise, leading to no annotation artifacts in the hypothesis. A state-of-the-art BERT-based model performs well on the CommitmentBank with 85% F1. However analysis of model behavior shows that the BERT models still do not capture the full complexity of pragmatic reasoning, nor encode some of the linguistic generalizations, highlighting room for improvement.

## 1 Introduction

Natural language inference (NLI), the task of identifying whether a hypothesis can be inferred from, contradicted by, or not related to a premise, has become one of the standard benchmark tasks for natural language understanding. NLI datasets, such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), are typically built by asking annotators to compose sentences based on premises extracted from corpora, so that the composed sentences stand in entailment/contradiction/neutral relationship to the premise. The hypotheses collected this way have been found to contain annotation artifacts – clues allowing systems to identify the relationship between a premise and a hypothesis without observing the premise. For instance, Gururangan et al. (2018) found that in SNLI and MultiNLI, negation is highly indicative of contradiction and generic nouns (e.g., *animal, something*) of entailment.

| | |
|---|---|
| *Premise:* A: Boy that's scary, isn't it. B: Oh, can you imagine, because it happens in the middle of the night, so you know, these parents didn't know the kid was gone until the kid is knocking on the door screaming, let me in.<br>*Hypothesis:* the kid was gone. | `Entailment (1.92)` |
| *Premise:* "Clever". Klug means "clever". Would you say that Abie was clever?<br>*Hypothesis:* Abie was clever. | `Neutral (0)` |
| *Premise:* GM confirmed it received U.S. antitrust clearance to boost its holding. Sansui Electric agreed to sell a 51% stake to Polly Peck of Britain for $110 million. Still, analysts said the accord doesn't suggest Japan is opening up to more foreign takeovers.<br>*Hypothesis:* Japan is opening up to more foreign takeovers. | `Contradiction (-1.2)` |

Table 1: Examples from the CommitmentBank, with NLI class and original annotation mean.

To address this issue, we recast the CommitmentBank (CB henceforth) (de Marneffe et al., 2019), an English dataset of speaker commitment/event factuality, for NLI.[1] The original CommitmentBank includes naturally occurring discourses annotated with speaker commitment towards the content of complements of clause-embedding verbs under entailment-canceling environments (negation, modal, question and conditional). CB does not suffer from the drawback of annotation artifacts in the hypotheses, since the hypotheses are the complement of a clause-embedding verb in the premise. It thus tests for

[1] The recast CommitmentBank is part of the SuperGLUE benchmark (Wang et al., 2019a), at https://super.gluebenchmark.com.

*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6086–6091,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

inferences involving a particular kind of syntactic construction and contains no annotation artifacts in the hypothesis, making it suitable to test for robust language understanding.

CB has many challenging aspects which are highlighted in various adversarial NLI datasets. It can be thought of as a variant of HANS (McCoy et al., 2019), which contains examples where the hypothesis is a subsequence or a constituent of the premise. It contains several phenomena in the "stress tests" (Naik et al., 2018) including word overlap, negation, and length mismatch. However these datasets are artificially constructed while CB data are naturally occurring.

Here we evaluate BERT, the state-of-the-art model in NLI, on CB. While BERT models achieve good performance with supervision from both CB and MultiNLI, they still struggle with items involving pragmatic reasoning and lag behind human performance. Experiments show that BERT does not use the linguistic generalizations for speaker commitment to make predictions, although BERT can learn them with direct supervision. CB is thus a useful benchmark for measuring progress on robust natural language understanding and specifically speaker commitment inferences.

## 2 The CommitmentBank

To study the linguistic correlates of speaker commitment in English, de Marneffe et al. (2019) introduced the CommitmentBank dataset.[2] It consists of naturally occurring English items with up to two sentences of preceding context and one target sentence, from three genres: newswire (Wall Street Journal), fiction (British National Corpus), and dialogue (Switchboard). The target sentences contain a clause-embedding verb (such as *think*) in an entailment-canceling environment (negation, modal, question, or conditional). Each item has at least 8 annotations indicating the extent to which the speaker of the sentences are committed to the truth of the embedded clause (+3/speaker is certain that it is true, 0/speaker is not certain about its truth, −3/speaker is certain that it is false).

### 2.1 Recast for NLI

For each item, we take the context and target sentence to be the premise, and the embedded clause in the target sentence to be the hypothesis.

| | Entailment | Neutral | Contradiction | Total |
|---|---|---|---|---|
| Train | 115 | 16 | 119 | 250 |
| Dev | 23 | 5 | 28 | 56 |
| Test | 113 | 16 | 121 | 250 |
| Total | 251 | 37 | 268 | 556 |

Table 2: Number of items with each gold label in each split. Split are from Wang et al. (2019a).

We identified a subset of the CommitmentBank with high annotator agreement, and assigned categorical labels (entailment/neutral/contradiction) to them according to their mean annotations in $[-3, 3]$. We label an item as entailment if at least 80% of its annotations are within $[1, 3]$, where the speaker is committed to the complement $p$, as neutral if within $[0]$, where the speaker is uncommitted toward $p$, as contradiction if within $[-3, -1]$ where the speaker is committed to $\neg p$. We discard the item if less than 80% of the annotations are within one of the three sub-ranges. Table 1 contains examples from CB with the original mean annotation and the gold NLI label. The number of items in each class is in Table 2.

### 2.2 Possible Annotation Artifacts

Since the hypotheses in CB are extracted from the premises instead of generated by annotators, we expect CB to contain less annotation artifacts compared to SNLI or MultiNLI.

**Length** Gururangan et al. (2018) found that entailed hypotheses in SNLI tend to be shorter and neutral ones longer.[3] The hypothesis length in the CB train set is distributed evenly across the three classes (mean length 8.5 tokens for entailment, 6.6 for neutral and 7.3 for contradiction).

**Lexical Features** Following Gururangan et al. (2018), we computed the PMI between each unigram/4-gram and class in the training set,[4] capturing the extent to which an expression is associated with each class. Table 3 gives the five unigrams and 4-grams with the highest PMI values. For hypotheses, there don't seem to be any discriminating expressions. In particular, negation words are not features for contradiction, in contrast to SNLI and MultiNLI. For premises, we find some expressions that are strongly associated with

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

[3] In SNLI, about half of the entailed hypotheses have token length less than 5, while a similar portion of the neutral hypothesis have length under 12 tokens.

[4] We applied add-one smoothing to the raw counts.

| **Hypothesis** | | | | | |
|---|---|---|---|---|---|
| Entailment | | Neutral | | Contradiction | |
| Mr | C ##rois ##set by | pretty | the cat was well | really | ’ s going to |
| been | the language was peeled | cat | at any given moment | people | a de ##ter ##rent |
| . | language was peeled down | moment | any given moment a | can | de ##ter ##rent to |
| had | no women are allowed | response | given moment a response | go | is going to be |
| - | women are allowed to | ban | moment a response of | parents | it was too long |
| **Premise** | | | | | |
| Entailment | | Neutral | | Contradiction | |
| g | might have known that | clever | Do you think that | mean | . B : I |
| herself | . You could say | Nicky | . Do you think | five | don ’ t think |
| perhaps | Mr . An ##tar | pressure | you know , what | care | B : Uh , |
| wrong | . An ##tar ’ | Base | ‘ ‘ I hope | guy | : I don ’ |
| notice | An ##tar ’ s | radio | ‘ I hope you | jury | , I mean , |

Table 3: Unigrams and 4-grams with top 5 PMI with each class for the hypotheses and premises. We used the BERT cased tokenizer, which marks split word pieces with ##. The 4-grams in the hypotheses listed here share the same PMI values with many other 4-grams in the same class (745 for entailment, 60 for neutral, 6 for contradiction).

a particular class. But most of these expressions align with linguistic generalizations about these particular constructions, as elaborated on below.

**Entailment** The most discriminating expressions for the entailment premises include modal operators *perhaps*, *could* and *might*. This is because 63 out of the 115 entailment items in the train set involve the modal environment. Factive verbs *notice* and *know* are also discriminating features of entailment, indicating that factive verbs tend to suggest the truths of their complement (Karttunen, 1971).

**Neutral** The most discriminating expressions for the neutral class include questions: *Do you think*. This is due to the fact that 10 out of 16 neutral items are under the question environment.

**Contradiction** For contradiction, the most discriminating expressions involve neg-raising constructions (*I don't think/know/believe that p*, where the speaker is committed to *p* being false), filler phrases *Uh* and *I mean*, and indicator of speakers in dialogues *B:*. These are all characteristic of the Switchboard genre, which makes up 80% of the contradictions in the training set.

## 3 Predicting NLI labels

We evaluate BERT, the state-of-the-art model for NLI, on CB.[5] The BERT model follows the standard practice for sentence-pair tasks as in Devlin et al. (2019). We concatenate the premise and the hypothesis with [SEP], prepend the sequence with [CLS], and feed the input to BERT. The representation for [CLS] is fed into a softmax layer for a three-way classification.

For all experiments, we used the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 1e-5, a batch size of 8, and fine-tuned with at most 10 epochs on each dataset. We fine-tuned BERT with three different sets of training data: CB only (CB$^B$), MultiNLI only (MNLI$^B$), and MultiNLI first then CB (MNLI+CB$^B$).[6] For comparison, we also included the models' performance on the MultiNLI dev set.

**Baselines** We included two baselines: a bag-of-words baseline (CBOW) in which each item is represented as the average of its tokens' GloVe (Pennington et al., 2014) vectors; a Heuristics baseline, only applicable to the CB dataset, which uses five rules based on the observations in Section 2.2: 1. items under modals are entailments, 2. neg-raising items of the form *I don't think/know/believe* are contradictions, 3. items with factive verbs are entailments, 4. items under negation are contradictions, 5. all other items are neutral.

**Human Performance** We included human performance on CB from Wang et al. (2019a) obtained by asking crowdworkers to re-annotate a part of the test set. The MultiNLI human performance accuracy is for the matched/mismatched test set from Nangia and Bowman (2019).

**Results** Table 4 shows the results. CBOW does not perform well on either datasets. On the CB

[5] We used `jiant` (Wang et al., 2019b) with the `bert_large_cased` model for all our experiments.

[6] Superscript $^B$ denotes BERT model tuned on the corresponding dataset to distinguish them from the actual datasets.

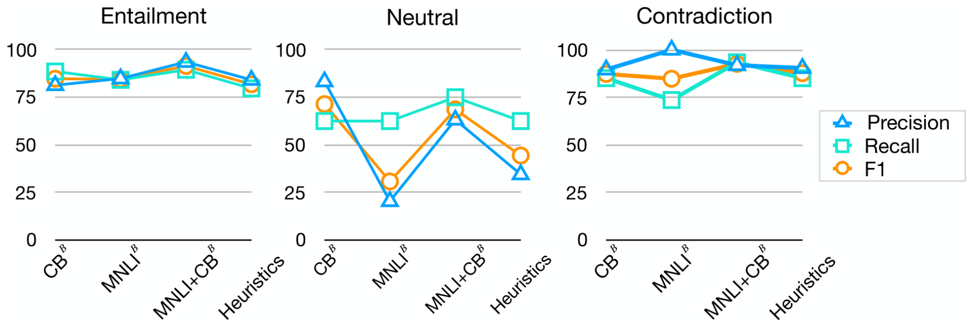

Figure 1: Precision, recall and F1 of each class on the CB test set for three BERT variants and Heuristics.

| | CB | | MultiNLI | |
|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 |
| CBOW | 69.2 | 47.6 | 71.2 | 71.2 |
| MNLI$^B$ | 77.6 | 66.7 | **83.6** | **83.6** |
| Heuristics | 81.2 | 71.3 | - | - |
| CB$^B$ | 85.2 | 81.2 | 41.1 | 30.6 |
| MNLI+CB$^B$ | **91.2** | **85.3** | 72.3 | 74.4 |
| Human | 98.9 | 95.8 | 92.0/92.8 | - |

Table 4: Performance on the CB test set and the MultiNLI dev set.

data, the heuristics based on linguistic generalizations is a strong baseline, performing better than MNLI$^B$. We gain a lot of performance with supervision from CB only (CB$^B$), which aligns with McCoy et al.'s observation that BERT performs very well when trained with in-domain data. The best results are obtained by MNLI+CB$^B$ on CB and by MNLI$^B$ on MultiNLI, but still lag behind human performance. While MNLI+CB$^B$ gives the best performance on CB, it does not perform well on MultiNLI. This is in line with Liu et al. (2019) who found that fine-tuning on datasets that test for a specific linguistic phenomenon decrease the performance on the original dataset.

## 4 Analysis

Figure 1 shows the precision, recall and F1 scores of each class on the CB test set for the three BERT variants and the Heuristics baseline. Heuristics performs similarly as CB$^B$ on all classes. Compared with CB$^B$, MNLI+CB$^B$ improves the overall performance of contradiction and the recall of neutral. MNLI$^B$ identifies contradictions with perfect precision but poor recall. All models do poorly on the neutral class, which has very few items in the dataset and no clear linguistic generalizations.

| Correct label with Heuristics? | Yes (203) | No (47) |
|---|---|---|
| CB$^B$ | 88.7 | 55.8 |
| MNLI$^B$ | 68.9 | 55.7 |
| MNLI+CB$^B$ | 89.4 | 68.5 |

Table 5: F1 scores of the three BERT models on the CB test set divided by whether Heuristics predicts the correct label (size of each subset in parentheses).

**Linguistic Generalizations and Beyond** 80% of the predictions of both CB$^B$ and MNLI+CB$^B$ are the same as the predictions of the Heuristics baseline, while it drops to 69.6% for MNLI$^B$. We divided the test set by whether Heuristics predicts the correct label. Table 5 reports the models' F1 scores on the two subsets. CB$^B$ and MNLI+CB$^B$ performances on the Yes-items are similar and both outperforming MNLI$^B$ (statistically significant improvements, McNemar's test, $p < 0.01$). There is no statistical difference between the models' performance for the No-items. There is thus a performance gap between items requiring more pragmatic reasoning in general (No-items) and items which can be correctly predicted by identifying certain structures (Yes-items), suggesting that there is still work to achieve robust language understanding. Table 6 shows some items on which MNLI+CB$^B$ still fails.

**Feature Probing** To investigate whether BERT actually learns the linguistic features from the Heuristics baseline and uses them to make NLI predictions, we trained two probing models to predict 1. whether the clause-embedding verb is factive and 2. the type of entailment-canceling environment. Following Tenney et al. (2019), we take the weighted sum of BERT layers (fine-tuned for NLI) to produce a pooled representation for each token. Unlike Tenney et al. (2019), in which the

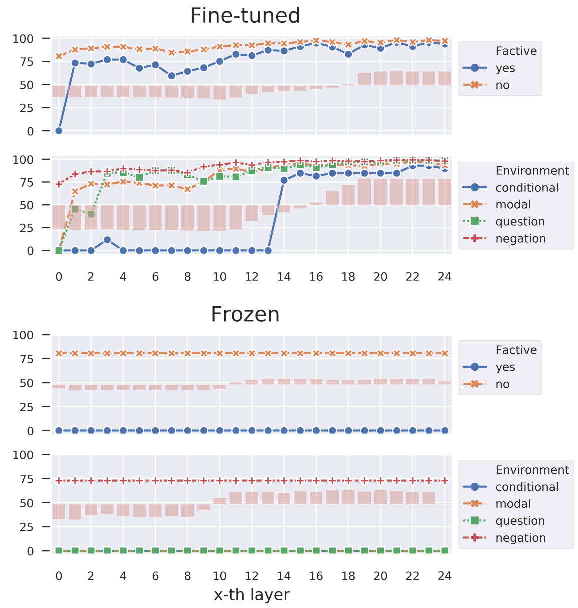

Figure 2: Results on the two probing tasks using parameters from MNLI+CB$^B$. Lineplots show model's F1 scores when using encoding layers up to the $x$-th layer. Barplots show the scalar mixing weights.

representations for the word tokens are used, we take the representation of the `[CLS]` token for each item and fed it into a MLP classifier to predict whether the discourse has certain features. We extracted the trained scalar mixing weights to see the importance of the different layers. For each layer $k$, we trained a series of classifiers using all previous layers up to $k$, to measure at which layer the feature can be correctly predicted.[7]

We did the above experiments in two settings: 1. fine-tune all BERT layers to learn the feature-specific representations, and 2. freeze BERT layers tuned for NLI and only train the probing classifier. The results are shown in Figure 2.

When fine-tuning BERT layers for each feature task, we see that performance increases as more layers are added. Factives, conditionals, and modals are correctly predicted at later layers than nonfactives and negation. For conditionals and modals, this might be due to the fact that they are rare in the dataset. Factives possibly require more contextual information in order to be learned: the scalar weights indicate that factivity is processed at higher layers than entailment-canceling environment. This is consistent with the language acquisition literature (Hacquard and Lidz, 2019) which suggests that rich syntactic/pragmatic information is required to learn the semantics of factive verbs.

*Premise:* B: Yeah, it's called VCX or something like that. Also called Delta Clipper, which is a decent name for something like that. A: Wow. Well, I don't know. you think you'd, uh, go up in space if you had a chance?
*Hypothesis:* speaker B would go up in space if he had a chance
H: `neutral` B: `contradiction` Gold: `neutral`

*Premise:* Those people... Not a one of them realized I was not human. They looked at me and they pretend I'm someone called Franz Kafka. Maybe they really thought I was Franz Kafka.
*Hypothesis:* he was Franz Kafka
H & B: `entailment` Gold: `contradiction`

Table 6: Items in the test set with predictions the by Heuristics baseline (H) and MNLI+CB$^B$ (B). The first one is correctly predicted by Heuristics, while the second one is not.

However, when we freeze the BERT parameters from MNLI+CB$^B$, the models always give the highest probability to negation environment and nonfactive verb, resulting in zero F1s on every other feature. The scalar mixing weights are smaller than the weights from the fine-tuned model. This suggests that, although BERT can learn these features with direct supervision, training BERT for NLI does not result in representations that encode these features: the model relies on other statistical clues to make decisions.

## 5 Conclusion

We introduce CB as a dataset for NLI, and show that it does not contain annotation artifacts in the hypotheses in contrast to previous NLI datasets. Our evaluation shows that despite the high F1 scores, BERT models have systematic error patterns, suggesting that they still do not capture the full complexity of human pragmatic reasoning. There is much room for improvement, and the CB dataset will be a useful testbed to assess models' progress on such reasoning.

## Acknowledgement

We thank Livy Real for suggesting to look at the CommitmentBank through the lens of NLI. We thank the members of the Clippers group at The Ohio State University for discussions and suggestions, as well as the reviewers for their feedback. This material is based upon work supported by the National Science Foundation under Grant No. 1845122.

[7] The code and data are available at `https://github.com/njjiang/jiant/tree/cb_emnlp19`.

## References

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)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Valentine Hacquard and Jeffrey Lidz. 2019. Children's attitude problems: Bootstrapping verb meaning from syntax and pragmatics. *Mind & Language*, 34(1):73–96.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Jason Phang, Edouard Grave, Najoung Kim, Phu Mon Htut, Thibault F'evry, Berlin Chen, Nikita Nangia, Haokun Liu, Anhad Mohananey, Shikha Bordia, Nicolas Patry, Ellie Pavlick, and Samuel R. Bowman. 2019b. `jiant` 1.1: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Adina Williams, Nikita Nangia, and Samuel R. 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*.