Re-Examining FactBank: Predicting the Author's Presentation of Factuality

John Murzaku[⋆◇], **Peter Zeng**[⋆], **Magdalena Markowska**[♣◇], **Owen Rambow**[♣◇]
[⋆] Department of Computer Science [♣] Department of Linguistics
[◇] Institute for Advanced Computational Science
Stony Brook University, Stony Brook, NY, USA
Corresponding Author: jmurzaku@cs.stonybrook.edu

Abstract

We present a corrected version of a previous projection of the FactBank data set. Previously published results on FactBank are no longer valid. We perform experiments on FactBank using multiple training paradigms, data smoothing techniques, and polarity classifiers. We argue that f-measure is an important alternative evaluation metric for factuality. We provide new state-of-the-art results for four corpora including FactBank. We perform an error analysis on Factbank combined with two similar corpora.

1 Introduction

The term *factuality*[1] refers to an author's presentation of a proposition (who-did-what-to-whom) as a fact, i.e., she is committed to the truth of the proposition. A lot of language use introduces propositions that are not presented as facts but as only possibly true, as a wish or as a hypothesis, or as someone else's belief. If we want to understand what an author is communicating, we need to distinguish these cases. Over the last 15 years, this question has received a lot of attention. Multiple corpora have been created, and these corpora have been used to explore machine learning architectures for factuality prediction. The machine learning studies often report results on all corpora, but these studies do not examine what the machine learning architecture can learn, nor how and why combining corpora can help. A notable exception is Jiang and de Marneffe (2021), who carefully analyze what exactly is learned from the CB corpus. In this paper, our goal is to determine how to combine corpora in order to maximize performance, and to understand why the specific combination of corpora works better than others. We choose a single resource to focus on so that we can gain insights by performing a careful study, and we choose FactBank (Saurí and Pustejovsky, 2009) because it is one of the first carefully constructed datasets for factuality prediction. We show how insights gained from working with FactBank can be used to improve performance on the CB corpus (de Marneffe et al., 2019).

There are three main contributions of this work:

(i) The FactBank data is complex. In order to facilitate NLP research, a machine learning-friendly projection from FactBank had been previously created and widely used. We correct an error in this projection (this data set will be made available). The error means that all recent results on FactBank are not valid.

(ii) We present new state-of-the-art results on the CB, FactBank, MV, and UW corpora.

(iii) We do an error analysis to show why some of these corpora perform better when combined.

This paper does not introduce new machine learning architectures; instead, we show that careful reexamination of the data can lead to improved performance without the necessity of introducing new, more complex architectures.

The paper is organized as follows. A survey of previous work is provided in Section 2. We summarize the FactBank representation of factuality in Section 3 and present our correction to the projection from FactBank in Section 4. In Section 5, we discuss metrics for evaluating factuality prediction. The redone experiments from (Jiang and de Marneffe, 2021) with our corrected projection are described in Section 6. We then report on machine learning experiments on three corpora: FactBank in Section 7, the CB corpus in Section 8, and the LDC corpus in Section 9. We conclude in Section 10.

[1] The notion of "factuality" is closely related to the notion of "belief" as used in cognitive science and AI; they differ only in the case of lying, where the author presents propositions as facts contrary to what she actually believes. See (Prabhakaran et al., 2015) for a fuller discussion.

Proceedings of the 29th International Conference on Computational Linguistics, pages 786–796
October 12–17, 2022.

2 Related Work

Corpora Many corpora explore the notion of *factuality* including: FactBank (Saurí and Pustejovsky, 2009), LU (Diab et al., 2009), UW (Lee et al., 2015), LDCCB (LDC) (Prabhakaran et al., 2015), MEANTIME (MT) (Minard et al., 2016), MegaVeridicality (MV) (White et al., 2018), UDS-IH2 (UD2) (Rudinger et al., 2018), CommitmentBank (CB) (de Marneffe et al., 2019), and RP (Ross and Pavlick, 2019). These corpora differ along several dimensions; we list dimensions which are salient for this paper.

(1) What type of data is used to build the corpus and is the data manipulated or not. E.g., MV selects only 6 syntactic frames and lexically “bleaches" them. In CB, only sentences with finite clausal complements are chosen for annotation. In addition, the matrix predicates must appear in the entailment cancelling environment, i.e. questions and negations preceded by a modal and/or in the antecedent of a conditional. In contrast, FactBank tags all events introduced in a corpus of complete, naturally occurring texts.

(2) What is the genre of the underlying texts? For those corpora which use naturally occurring texts, FactBank and UW use newswire exclusively. MT uses Wiki articles. RP uses data from textual inference corpora. CB uses newswire, fiction and dialog. UD2 uses weblogs, newsgroups, email, reviews, and question-answer corpora. LDC is exclusively discussion forum threads.

(3) The definition of an annotatable event. E.g., in MV only past events are taken into consideration; in UD2, both past and present events; UW, FactBank, CB and LDCCB consider also future events.

(4) Who are the annotators? FactBank and LDCCB used trained annotators, while CB and UW argue that crowd-sourced judgements collected from naive annotators are as (or more) valuable.

(5) The annotation scale. Annotations can be numerical values (often derived from averaging naive annotators’ judgments), typically [-3,3] (CB, UW, RP or UD2), lexically represented values, such as [yes, maybe or maybe not, no] in MV, or categorial labels (FactBank and LDCCB). We discuss these categorial labels in more detail in Section 3, and give examples.

Stanovsky et al. (2017) unify the representation across datasets up to 2017 by mapping the discrete annotations of factuality in FactBank and MT onto the continuous scale used in UW. Furthermore, they also remove the FactBank non-author perspective annotations since none of the other corpora include such annotations (FactBank also annotates the beliefs of agents mentioned in a sentence, according to the author). This process will be discussed in more details in Section 4.

Experiments Early work on event factuality prediction used rule-based systems; for example, Nairn et al. (2006) propose a recursive polarity propagation algorithm which uses implication signatures from clause-embedding verbs. Lotan et al. (2013) predict factuality using implication signatures combined with lexical and dependency tree features.

Early machine learning work on event factuality prediction consists of SVMs or other supervised learning approaches. Diab et al. (2009) and Prabhakaran et al. (2010) use SVMs and CRFs along with lexical and dependency tree features for predicting author belief, evaluating on f-measure. Lee et al. (2015) also use an SVM along with lexical and dependency tree features on the UW corpus which they created, and evaluate on Pearson correlation and mean average error (MAE), as does all following work. Stanovsky et al. (2017) use SVMs combined with the output of the system of Lotan et al. (2013), and evaluate on Factbank, UW, and Meantime. Rudinger et al. (2018) use bidirectional LSTMs with tree or linear architectures and multiple task-specific training setups. Building on that work, Veyseh et al. (2019) use BERT sentence representations combined with a graph convolutional network that leverages the semantic and syntactic structure of the sentence, evaluating on Factbank, UW, Meantime, and UD2. At the time of publication, their system produced state-of-the-art results for Pearson correlation on FactBank, UW, Meantime, UD2, and state-of-the-art results for MAE on FactBank, Meantime, and UD2, with Stanovsky et al. (2017) still having the lowest MAE for UW. We discuss Jiang and de Marneffe (2021) in more detail in Section 6.1.

Our work differs from the related work by offering two salient contributions: first, we analyze specific corpus combinations to help improve on a specific corpus (FactBank) instead of focusing on improving all corpora. Second, we perform error analyses on corpus combinations to determine why and where factuality corpora can help each other or why factuality corpora can be incompatible.

3 Representation of Factuality

As discussed in Section 2, some corpora represent the factuality judgments using numbers in an interval, while others use categorial labels with a defined meaning. The corpus creators determine that there are several distinct categories that annotators can identify. In the examples (all from FactBank, some simplified), the head of the phrase which presents the evaluated situation is bolded.

Certain (FactBank label CT): the author commits to the truth of the presented situation. Note that the commitment is independent of tense.

(1) A lawsuit in Germany will **seek** a criminal prosecution of the Defense Secretary.

Probable (FactBank label PR): the author presents the situation as probable.

(2) Saddam appeared to **accept** the treaty.

Possible (FactBank label PS): the author presents the situation as possible.

(3) He won't be under control until he is **committed** to an institution.

Fully underspecified (FactBank label UU): The source does not know what is the factual status of the event, or does not commit to it.

(4) The minister denied the kingdom had **notified** its customers.

In addition, FactBank annotates polarity on types of author belief, i.e., whether the author presents the situation as factual, or the absence of the situation as factual. Polarity is only added to the non-UU values, resulting in a label set containing seven labels, CT-, PR- PS-, UU, PS+, PR+, CT+. We discuss our mapping between the categorical labels and numerical labels in Section 7 (see Table 4).

4 Corrected FactBank Dataset

Label Correction Stanovsky et al. (2017) developed a projection from the complex FactBank corpus to a CoNLL-formatted file that includes only factuality judgments by the author, enabling an annotation with a single value at the word level. This data set has been extensively used in NLP experiments (see Section 2). We have found that there is a systematic error in this FactBank data set projection. Consider (4) above. FactBank annotates the denial event as seen as factual by the author (CT+). The notification event is annotated twice: according to the author, the minister sees it as certainly false (CT-). The author herself does not express her view of the factuality and her perspective is labeled UU, as explained in Section 3. The old projection of the FactBank data set incorrectly picks up the CT- label. We correct the projection by sticking to the author's perspective and supply the UU label. Table 2 shows the shift in label distribution percent towards UU as a result of our correction.

Article Split Correction Furthermore, Stanovsky et al. (2017) do not split by article, meaning that data from the same article could appear in both the training and test sets. We re-split the data, this time assigning all annotations from a single article to the same split, until we approximate a standard ratio of 70-20-10 for train-dev-test. With the new projection of FactBank and article-based split, we found that when training and testing on FactBank, there is a 2% decrease in Pearson correlation and a 10% increase in MAE compared to the previous projection of FactBank without the article-based split.

We will make the correct projection of the FactBank data set available, see Appendix A for details.

5 Evaluation of Factuality Prediction

To date, the category labels in FactBank have been translated to numbers for training and testing (from -3 for CT- to +3 for CT+). A regression head predicts a number, and the results are evaluated with the Pearson correlation r between the predicted and gold numbers. The publications to date also provide MAE (which need not correlate with correlation). We propose an additional evaluation, namely f-measure on the categories and macro averaged f-measure.

Usually in NLP, there is no single correct evaluation metric: the best evaluation metric to use depends on the downstream use we want to make of the module we are evaluating. In some applications, we need to know the specific level of commitment of the writer. For example, in argumentation analysis we need to detect claims to which the writer is fully committed, and when analyzing hedging as a marker of politeness or power structure, we need to identify the PR/PS family. But the numerical evaluation makes the same difference between PR+ and PS+ on the one hand, and PR+ and CT+ or PS+ and UU on the other hand. F-measure clearly separates these cases. In this paper, we show results using correlation, MAE, and the f-measures (macro-averaged F1 and per-label F1).

	FB-Old (NS)		FB-New (NS)		FB-New (S)		Previous SOTA	
Test Set	**r ↑**	**MAE↓**	**r↑**	**MAE↓**	**r↑**	**MAE↓**	**r↑**	**MAE↓**
CB	0.890	0.617	0.908	0.613	0.906	0.561	0.890	0.617
RP	0.870	0.608	0.856	0.630	0.861	0.642	0.870	0.608
MV	0.857	0.533	0.867	0.498	0.886	0.483	0.876	0.501
MT	0.491	0.319	0.456	0.311	0.553	0.281	0.702	0.204
UW	0.865	0.351	0.879	0.366	0.874	0.348	0.868	0.349
UDSIH2	0.853	0.766	0.855	0.763	0.857	0.758	0.909	0.726
FB-New	-	-	0.858	0.359	0.866	0.330	-	-

Table 1: Results on multi-task training with no smoothing (NS), smoothing (S), and the new projection of FactBank (FB-New); a dark shaded cell indicates the best published result to date in this table; light shading means an improvement over FB-Old (NS). Results for all corpora except RP show improvement. All other state of the art results for non-shaded cells are held by Jiang and de Marneffe (2021) or Veyseh et al. (2019).

	CT+	PR+	PS+	UU	PS-	PR-	CT-
Old	75.3	3.1	2.2	14.5	0.1	0.6	4.2
Corrected	57.1	1.1	1.1	38.3	0.1	0.1	2.2

Table 2: Distribution of each FactBank annotation label in the old and corrected CoNLL-formatted data set

6 Evaluation on All Corpora

6.1 Redoing (Jiang and de Marneffe, 2021)

Jiang and de Marneffe (2021) provide state-of-the-art results for many of the corpora discussed in Section 2 using a simple architecture. In this section, we redo their experiments with the updated FactBank data set. Specifically, we redo the multi-task learning experiments using the same underlying architecture (the SelfAttentiveSpanExtractor developed by Gardner et al. (2018)) and the same training parameters as Jiang and de Marneffe (2021). This training paradigm allows all corpora to share the same BERT parameters, but with each corpus having a regression head with corpus-specific parameters. The authors find that fine-tuning on BERT-large performs best. However, with the corrected FactBank dataset, we find that RoBERTa-large outperforms BERT-large, and we therefore use it for our experiments. For each experiment, we do three runs with different seeds and report the average for Pearson correlation and MAE, and for most experiments we also provide the standard deviation.

Table 1 shows results with the old and corrected data sets. We replicate the results from Jiang and de Marneffe (2021) using the faulty dataset used in previous experiments (columns **FB-Old (NS)**). For our corrected dataset (columns **FB-New (NS)**), results for all test sets other than RP improve, presumably because the FactBank data is now more in line with the other corpora. Note that none of the results in this paper for FactBank are comparable to any previously published results because of the errors in the FactBank data set used to date (see Section 4), which means that the FactBank test set has also changed.

6.2 Addressing Imbalances in Corpora

One problem with all corpora in this study, including FactBank, is the inequality in the label distribution, with a majority of CT+ (3.0) and UU (0.0) as shown in Table 2 (this effect holds in all corpora, including those with purely numerical labels). We address this issue by performing label distribution smoothing and modifying the loss function to a weighted SmoothL1 loss.

Yang et al. (2021) provide methods to address class imbalance problems in a continuous setting using label distribution smoothing and feature distribution smoothing, which directly applies to our regression task. We apply their method of label distribution smoothing (LDS) to our datasets by using kernel density estimation to learn the effective label density in our dataset. We then re-weight the SmoothL1 loss function by multiplying it by the inverse of the effective label densities learned. This method improves on all of our tasks, so we perform all of our regression experiments using LDS unless otherwise noted. Results for redoing Jiang and de Marneffe (2021) experiments with LDS are shown in Table 1 (columns **FB-New (S)**). We see an improvement compared to no LDS (columns **FB-New (NS)**) for all corpora on both metrics (correlation and MAE), except for correlation on CB and UW.

Single	r↑	MAE↓	r↑	MAE↓
FB	0.872±0.002	0.276±0.004	0.872±0.002	0.276±0.004
	Shared		**MTL**	
FB+CB	0.876±0.011	0.293±0.029	0.873±0.008	0.292±0.015
FB+MV	0.874±0.000	0.289±0.028	0.885±0.005	0.274±0.034
FB+RP	0.871±0.005	0.293±0.013	0.879±0.005	0.311±0.009
FB+MT	0.864±0.010	0.284±0.011	0.875±0.007	0.334±0.039
FB+UD2	0.818±0.023	0.386±0.037	0.867±0.007	0.360±0.009
FB+LDC	0.802±0.075	0.343±0.066	0.868±0.010	0.329±0.039
FB+UW	0.741±0.034	0.717±0.080	0.873±0.007	0.289±0.023
FB+CB+MV	0.881±0.001	0.278±0.005	-	-
FB+CB+MV+RP	0.873±0.007	0.316±0.005	-	-
FB+MV+RP	-	-	0.879±0.011	0.305±0.022

Table 3: Results for our regression experiments on FactBank. Single results show FactBank trained and tested on itself. The Shared and MTL columns show Pearson r and MAE on the two training paradigms respectively. A shaded cell indicates the best performing combination; light means only a slight improvement.

7 Experiments on FactBank

In this section, we evaluate exclusively on FactBank (using our new FactBank dataset described in Section 4).

Our goal is to provide the best results on FactBank. For all experiments reported in this section, we follow the same setup as Section 6, except that we train two models per setup, one for regression which we evaluate numerically (r and MAE), and one for classification which we evaluate using F1. In classification, we collapse FactBank labels PR and PS (probable and possible), as they are rare, and their distinction is less important. Table 4 shows the mappings of the continuous labels to discrete labels with the corresponding factuality values (in FactBank terms). Following Jiang and de Marneffe (2021), we perform our training in two ways: **Shared**, where we combine the corpora's training data together and test on Factbank; and multi-task learning (**MTL**), where corpora share the same RoBERTa-large parameters, but we have a corpus-specific regression or classification head for FactBank. All experiments are performed three times with different seeds, and we report the average and standard deviation. For further information on the experimental setup, see Appendix B.

Range	Label
[-3.0, -2.5]	CT-
(-2.5, -0.5]	PR-
(-0.5, 0.5)	UU
[0.5, 2.5)	PR+
[2.5, 3.0]	CT+

Table 4: Mappings for our classification model.

7.1 Regression Experiments

We perform corpus combinations of FactBank with each of the other corpora, and evaluate on Pearson correlation and MAE. We then do a greedy search with the top performing corpus combinations: we take our best performing system and add the corpus which performs next best in the 2-way combination with FactBank. If adding another corpus does not yield an improvement, we stop our search. All results for both the **Shared** and **MTL** experiments are shown in Table 3.

Shared: FactBank combined with CB and FactBank combined with MV yield improvements in correlation, but not in MAE. In our greedy search, we combine FactBank with both CB and MV. FactBank combined with CB and MV performs the best on correlation achieving a result of 0.881. However, this corpus combination results in a slightly higher MAE compared to baseline FactBank only. We then add RP, but FB+CB+MV+RP performs worse and we end our greedy search. The worst performing corpus combination on FactBank is FB+UW, resulting in the lowest correlation and also a very high MAE. This is because of article overlap in FactBank and UW with different annotation labels as shown in Lee et al. (2015), leading to a divergence.

MTL: The top performing result on FactBank is with MTL on FB+MV, with a Pearson correlation of 0.885 and MAE of 0.274. Each corpus combination besides FB+UD2 results in an improvement over baseline FactBank. MTL improves on the worst performing corpus combinations in the Shared paradigm (FB+UW and FB+UD2): the FactBank specific corpus head in the model is optimized for FactBank, and therefore addresses the lack of performance caused by treating all corpora as one. We perform a greedy search by adding RP to our top performing MTL combination of FB+MV. The correlation is higher than FactBank Single, but not higher than the FB+MV combination, so we stop our greedy search.

7.2 Classification Experiments

We perform all classification experiments with the same hyperparameters and training architecture as our regression experiments, but with a classification head instead of a regression head. We mention more details about this in Appendix B. Even after collapsing PR±/PS± labels, PR±/PS± are a minority class in our classification experiments. We address the label imbalance in our classification models by using focal loss (Lin et al., 2017), which has been shown to perform well on imbalanced classification tasks compared to cross-entropy loss. Using the focal loss function allows our model to focus on the harder to classify PR±/PS± examples. All results for both the **Shared** and **MTL** experiments are shown in Table 5.

Shared: The largest increase is achieved by combining FactBank with CB, with a major boost in macro-average and in the per label F1 of the minority labels PR± because CB introduces many new PR± labels. The only other corpus that achieves an increase on FactBank is RP, which specifically helps in the UU and CT+ labels. All other corpora do not help, notably UD2 and UW, which result in a massive decrease. Some per-label F1s are 0 for FB+LDC. The LDC corpus does not contain polarity, and therefore has no labels in the CT- and PR- categories. We perform a greedy search on our top performing corpora combinations, adding RP to FB+CB, but do not improve, so we stop our greedy search.

MTL: Again, CB helps the most, with the highest macro-average on FactBank in our 5-way system, specifically helping with the minority classes PR- and PR+. MT, UW, and MV also provide a boost. The previous poor performance of UW is fixed by training in a MTL setting. All other corpora decrease performance on FactBank. UD2 performs poorly, but slightly better than in the shared setting. We perform a greedy search and stop after adding MT to yield FB+CB+MT, which does not improve on our top combination of FB+CB.

7.3 End-to-end Evaluation

One advantage of the F1 evaluation is that we can provide a single end-to-end evaluation on data without gold heads. All experiments reported in the literature using r and MAE assume a gold head, since otherwise these measures cannot be computed. We introduce a "Not-Head" tag (O) for all words that are not heads and train a model on all words in the data set, not just the heads. We repeat the experiments with the same architecture and setup as the previous classification experiments. The results are shown in Table 6, and as expected, the F1s for each class decrease, but not dramatically. The F1 for detecting that a word is a head is 0.888; whether a noun is an event is context-dependent (e.g., *construction*) and can be hard to determine. Since the head-identification task is not trivial, we argue that an evaluation on F1 is therefore important and can offer broader insights into the factuality-prediction task.

7.4 Factoring Polarity

Polarity is often expressed independently of the degree of factuality, as illustrated in this constructed example: *Sudeep {probably/maybe/∅} {came/did not come} to dinner.* Here, the first set of curly brackets lists three options for factuality, and the second set of curly brackets determines polarity. All six combinations are plausible sentences, determining six different FactBank labels for the coming event. Of course, some lexical items precisely encode a combination of polarity and factivity level (such as *deny* in (4)).

We train a polarity classifier on the same training data, but this time we label all negative data (CT-, PR-, PS-) as a new label NEG, the neutral data remains UU, and the positive data (PS+, PR+, CT+) as POS. The classifier performs with per-label F1-measure of 0.907 (Neg), 0.824 (UU), and 0.940 (Pos). We then create a combined system with our polarity classifier and our 5-way classification system where the polarity classifier assigns the polarity of the head, and the classification system predicts the strength. Our results for this system on FactBank are shown in Table 7.

On our FactBank-only system, the polarity classifier results in a 28% error reduction on macro-average, while also stabilizing results by lower-

Single	Macro-F1	CT-	PR-	UU	PR+	CT+	Macro-F1	CT-	PR-	UU	PR+	CT+
FB	0.701±0.076	0.863	0.222	0.893	0.593	0.935	0.701±0.076	0.863	0.222	0.893	0.593	0.935
	Shared						MTL					
FB+CB	0.790±0.053	0.863	0.611	0.885	0.662	0.932	0.800±0.017	0.851	0.667	0.880	0.677	0.930
FB+RP	0.717±0.089	0.843	0.444	0.889	0.475	0.936	0.655±0.009	0.874	0.000	0.888	0.581	0.933
FB+MT	0.703±0.088	0.852	0.222	0.894	0.614	0.936	0.749±0.088	0.861	0.389	0.895	0.668	0.937
FB+MV	0.679±0.1200	0.833	0.222	0.883	0.530	0.931	0.705±0.064	0.818	0.222	0.880	0.680	0.930
FB+LDC	0.631±0.027	0.759	0.000	0.887	0.580	0.930	0.664±0.010	0.847	0.000	0.886	0.657	0.932
FB+UD2	0.566±0.026	0.690	0.060	0.864	0.292	0.923	0.624±0.000	0.846	0.000	0.883	0.462	0.929
FB+UW	0.546±0.015	0.797	0.122	0.705	0.212	0.894	0.744±0.084	0.866	0.444	0.879	0.601	0.930
FB+CB+MT	-	-	-	-	-	-	0.693±0.051	0.867	0.222	0.891	0.551	0.935
FB+CB+RP	0.704±0.076	0.855	0.355	0.888	0.489	0.934	-	-	-	-	-	-

Table 5: Results for our classification experiments on FactBank. The topmost results show FactBank trained and tested on itself as a baseline. The Shared and MTL columns show F1 and per-label F1 on the two training paradigms respectively. A shaded cell indicates the best performing combination; light means only a slight improvement.

Macro-F1	CT-	PR-	UU	PR+	CT+	O
0.727	0.519	0.667	0.735	0.714	0.767	0.961

Table 6: Results for our end-to-end classification system on FactBank+CB.

Train	NoPC Macro-F1	PC Macro-F1	ER%
FB	0.701±0.077	0.786±0.003	28
FB+LDC	0.665±0.011	0.776±0.017	33

Table 7: Results on macro-average for our polarity classifier jointly combined with our 5-way classifier for without polarity (NoPC Macro-F1), with polarity (PC Macro-F1), and error reduction (ER).

ing standard deviation. The polarity classifier also helps other combinations. Our highest error reduction is for FB+LDC, presumably because LDC does not contain polarity annotations. We achieve a 33% error reduction in MTL training and 29% error reduction in Shared training; on regression, we also achieve an error reduction of 33%. There is also a moderate error reduction in classification for FB+UW (19% in Shared training and 11% in MTL training) and FB+UD2 (16% in Shared training and 10% in MTL training). For MV, we only get an error reduction in the MTL training setting of 10%. However, we do not obtain error reductions using this technique for our best performing combinations.

7.5 Why Does CB Help FactBank?

FactBank is all newswire, and CB also contains newswire. This does not fully explain why CB helps with FactBank. To further examine the issue, we consider two models, the model trained only on FactBank (FB) and the MTL classification model trained on FactBank and CB (FB+CB). We perform an error analysis on the data points in the FactBank dev set on which the two models make different predictions. Our goal is to determine how using the CB corpus helps. The results are in Table 8. We use the following categories; the first five are morpho-syntactic. **Noun** means that the target is a noun designating an event not tagged UU in FactBank; **Noun-UU** is a noun which is tagged UU. **Main** refers to heads that are main clause verbs or verbs in adjunct clauses to the main verb; these are typically easy cases. **Embedded** refers to targets which are in complement clauses below a main clause verb; the factuality status is typically determined or strongly affected by the main clause verb. **Relclause** refers to heads which are in relative clauses. **Hypo** are hypothetical situations. **Idiom** groups together various cases of idiomatic language use, either multiword expressions or idiomatic usages of lexical items. **Misc** groups together various other syntactic and semantic special cases.

We observe that 10% of all errors are gold errors (i.e. errors in the original annotation), and furthermore, that FB makes more errors in total than FB+CB, since FB+CB performs better. (The percentages do not sum to 100% in each row since some errors are made by both models.) Recall that CB annotates only verbs in embedded clauses. We therefore expect the FB+CB model to perform poorly on main clause verbs (**Main**) and nouns, which is borne out. The exception is nouns labeled UU, since the FB+CB model appears to label most

Type	Nb	FB	FB+CB	Gold
Noun	23	22%	70%	9%
Noun-UU	4	100%	0%	0%
Main	3	0%	100%	0%
Embedded	17	65%	24%	18%
Rel clause	9	78%	22%	0%
Hypo	4	100%	0%	0%
Idiom	6	67%	17%	17%
Misc	16	69%	44%	13%
Total	82	56%	40%	10%

Table 8: Error analysis on differences between FactBank-only model and FB+CB model; percentages refer to portion of all the errors of that type made by the two systems and Gold; percentages in one row can sum to more than 100% because the same error can be made by both systems, or by a system and Gold

nouns as UU by default, thus getting UU nouns correct by accident. The **Misc** error category is balanced between the two models, and in all other models FB+CB performs better. The biggest such category is **Embedded**, which is precisely what CB annotates, and for which the FB+CB model has far fewer errors than the FactBank model alone (24% of 17 errors against 65%). The error analysis thus shows that by adding CB to the multi-task training, the resulting model has learned what CB is designed to provide information on (embedded verbs), but suffers from the lack of representative data distribution in CB and increases errors for categories that CB does not not annotate (main verbs, nouns).

The results of the error analysis suggest another type of system: we use the FactBank-only system for noun heads, and the FB+CB system for verb heads (the number of main verb errors is small, so we do not worry about syntax). We implement this system using the Spacy POS-tagger (Honnibal and Montani, 2017), and using the previously trained models. If the Spacy POS-tagger tags a head as a noun, we use the FactBank-only system; otherwise we use the FB+CB system. The results are shown in Table 9. We see that this strategy provides us with the best result for 5-way classification, decreasing macro-average error by 12% and improving on the CT-, UU, and PR+ labels. We also note that this system has the smallest standard deviation among all models that perform at baseline or above, suggesting that the system is consistent in its behavior. We also perform this method on FB+CB in regression, and obtain an increase in Pearson correlation over the results in Table 3 from 0.876 to 0.888 (error reduction of 10%) and a slight decrease in MAE from 0.293 to 0.286 (error reduction of 2%).

	Macro-F1	CT-	PR-	UU	PR+	CT+
FB+CB	0.800$_{\pm 0.017}$	0.851	0.667	0.880	0.677	0.930
FB+CB'	0.824$_{\pm 0.006}$	0.873	0.667	0.887	0.765	0.930

Table 9: Results on FactBank using FB+CB as a baseline and FB+CB with a POS switch (FB+CB'). We show macro-average (Macro-F1) and per-label F1 performance. Shaded cells indicate improvements.

7.6 Summary on Experimental Findings

Regression Experiments Our first insight is that the new projection of FactBank helps on all corpora as shown in Table 1 because this projection makes more sense for a system to learn. On our FactBank focused-experiments, we find that the CB corpus helps FactBank the most in Shared, while the MV corpus helps FactBank the most in MTL as shown in Table 3.

Classification Experiments We find that CB helps FactBank in both the Shared and the MTL setting, outperforming all other corpus combinations as shown in Table 5. Furthermore, we find that our best results (FB+CB) can be further improved by using a POS-based system and we see improvement on all metrics as shown in Table 9. Finally, we find that an end-to-end system has predictably lower performance on f-measure as shown in Table 6, while offering the advantage of not assuming gold heads.

8 Testing on CB

We have seen that the corpora that most help FactBank are CB and MV. Can FactBank and MV help CB? In Section 6, we trained on all corpora (following the lead of Jiang and de Marneffe (2021)). Here, we train on only CB, FactBank, and MV, and evaluate on CB. The results are shown in Table 10 for CB using multi-task learning; similar experiments testing on MV did not yield improvements over our new state-of-the-art results in Table 1. When training on FB+CB, we do not improve on Pearson correlation but slightly

Train	r	MAE
FB+CB	0.885	0.602
FB+CB+MV	0.913	0.536

Table 10: MTL training on CB, FactBank, testing on CB test set

improve on MAE compared to previous CB results. However, when training on FB+CB+MV MTL, we achieve state-of-the-art results on CB in both Pearson correlation and MAE, providing a 21% error reduction on both measures (coincidentally) over the state-of-the-art prior to this paper.

9 Testing on LDC

Experiments We perform a separate set of experiments with the LDCCB corpus (LDC) (Prabhakaran et al., 2015) and FactBank on LDC. We choose the LDC corpus because it is similar to FactBank with reference to the annotation goals and the use of expert annotators. One major difference, however, is genre. LDC consists of discussion forum posts with many typos and fragmentary language, while FactBank consists entirely of newswire. We carefully examine this combination and show why some factuality corpora, even if they have similar annotation goals and annotators, may be incompatible. Using the same system and experimental setup as Appendix B, we perform two experiments: first, we train on LDC and test on LDC, and second, we train on FactBank and test on LDC. Table 11 shows results for these two experiments respectively. We see that FactBank performs very poorly on LDC, highlighting a potential incompatibility and mirroring our results training on FB+LDC in Table 3 and testing on FactBank, where LDC was among the worst performers.

Train	**r**	**MAE**
LDC	0.822	0.361
FB	0.616	0.630

Table 11: Results on LDC corpus test-set trained on LDC and trained on FactBank (FB).

Error Analysis We perform an error analysis on the application of the FactBank-only model on LDC, choosing 100 errors randomly. We list the percentage of errors in each category. We use the same categories as in Section 7.5, though there are no errors on nouns, and we add the category **Unclear** which includes fragments, typos, grammar and spelling errors which are the effect of unedited text. First, the gold standard errors are strikingly high: out of 100 examples, 5 heads were labeled in error in gold, and 45 of the remaining examples are mislabeled in the gold standard. We assume that this percentage is not representative of the corpus, since these are the difficult cases, but it also may be that annotation is harder on informal domains. As expected for a model trained on edited newswire, **Unclear** is the top error category. Ignoring **Misc**, the next biggest error categories are **Embedded** and **Hypo**. We already saw this weakness in the FactBank-only model on hypotheticals in Table 8. In conclusion, the error analysis shows the importance of genre (**Unclear** errors), and the continued weakness of FactBank-trained models for hypotheticals and embedded clauses.

Type	%
Main	3
Embedded	12
Hypo	11
Unclear	13
Misc	11
Gold	50
Total number	100

Table 12: Error analysis on FactBank system prediction on LDC

10 Conclusion

After correcting an error in a widely used data set derived from FactBank, we report new best results on four corpora: FactBank (Table 3), CB (Table 10), MegaVeridicality (Table 1), and UW (Table 1). We also provide f-measure evaluation, and extend this to a true end-to-end evaluation, the first in the literature. Finally, we show that by combining compatible corpora (FactBank, CB, MV), we can achieve improvements in performance on FactBank and CB, and that the improvements on FactBank are precisely as expected given how CB was created.

Given the targeted help CB can provide on FactBank predictions for embedded clauses, and given the current weakness on hypotheticals in FactBank, we suggest a new targeted annotation of factivity on sentences in hypothetical contexts.

Acknowledgments

We thank: Bonnie Dorr and Brodie Mather for encouragement and interest; the three anonymous reviewers for their insightful and useful comments; Nan-Jiang Jiang for help with the Jiant system and the datasets; Amittai Aviram and Tyler Osborne for help with the data representation. This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contracts No. HR01121C0186, No. HR001120C0037, and PR No. HR0011154158. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

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, Lisbon, Portugal. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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, Atlanta, Georgia. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

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)*.

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

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China. Coling 2010 Organizing Committee.

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.

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.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *ACL*.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. *arXiv preprint arXiv:1907.03227*.

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, 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.

Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. 2021. Delving into deep imbalanced regression. In *International Conference on Machine Learning (ICML)*.

A Distribution of New Data Set

We intend to distribute the corrected FactBank data set. We have included the training portion in this submission for reviewers to inspect, but we cannot distribute it for copyright reasons. Instead, we will provide a Python script which will produce the files submitted with this paper from the original FactBank files. These files can be obtained by researchers from the Linguistic Data Consortium, catalog number LDC2009T23. The entire corpus contains 9,740 annotated data points, split by article.

B Details on Experiments

We used a standard fine-tuning approach on existing BERT-large and RoBERTa large models with 333,843,458 and 355,623,938 paramaters respectively. For computing, we used our employer's GPU cluster. Compute jobs were typically about 25 minutes on average and ran on a single Tesla V100-SXM2 GPU. We did not do any hyperparameter search or hyperparameter tuning. We followed the same training parameters as Jiang and de Marneffe (2021), where we fine-tuned our model for at most 20 epochs with a learning rate of 1e-5. Early stopping was used if the difference between Pearson r and MAE did not increase, or if macro F1 did not increase. All metrics for experiments were averaged over three runs using fixed seeds (7, 21, and 42) which we will share with our code. We have also noted where testing for statistical significance of results was performed and have provided standard deviations for our results. To fine-tune the models and run experiments, we used the. jiant-v1-legacy library (Wang et al., 2019) and the implementation of Jiang and de Marneffe (2021) which uses jiant-v1-legacy. We added the classification module for the Jiang implementation and will make that available. All evaluation was performed by the jiant-v1-legacy library. All pre-processing scripts will be made available.