Towards Generative Event Factuality Prediction

John Murzaku[♣◇△], **Tyler Osborne**[♠], **Amittai Aviram**[♠], **Owen Rambow**[♣◇□]
[△] Department of Computer Science [□] Department of Linguistics
[◇] Institute for Advanced Computational Science
[♣]Stony Brook University, Stony Brook, NY, USA
[♠]Department of Computer Science, Boston University, Boston, MA, USA
Corresponding Author: jmurzaku@cs.stonybrook.edu

Abstract

We present a novel end-to-end generative task and system for predicting event factuality holders, targets, and their associated factuality values. We perform the first experiments using all sources and targets of factuality statements from the FactBank corpus. We perform multi-task learning with other tasks and event-factuality corpora to improve on the FactBank source and target task. We argue that careful domain specific target text output formatting in generative systems is important and verify this with multiple experiments on target text output structure. We redo previous state-of-the-art author-only event factuality experiments and also offer insights towards a generative paradigm for the author-only event factuality prediction task.

1 Introduction

The term *factuality* refers to the author's or speaker's presentation of an event as factual, i.e. as an event that has happened, is happening, or will happen. Often times, an author does not only talk about what they believe is factual, but also about what others believe is factual. Thus, when a speaker presents an event, they communicate their view of the factuality of the event, and they can also at the same time attribute a factuality judgment about the same event to another source. Over the past 15 years, the task of event factuality prediction has received a lot of attention, but only in predicting the factuality of an event according to the author's presentation. Multiple corpora have been created alongside multiple machine learning architectures which solely focus on predicting the author's presentation of factuality.

An exception is the FactBank corpus (Saurí and Pustejovsky, 2009), which not only annotates the author's presentation of factuality, but also annotates the nested sources assigning factuality values to events in text. In this paper, our goal is to predict the presentation of factuality of the nested sources mentioned in a text alongside their target events. We choose the FactBank corpus (Saurí and Pustejovsky, 2009) as it is the only corpus annotating nested source factuality and it is carefully annotated and constructed. We attempt combinations with other corpora, namely author-only event factuality corpora and source and target cognitive state corpora, to improve on predicting nested source and target factuality. We perform all of these experiments with a novel generative approach and create a new version of the event factuality prediction task.

There are four main contributions of this work:

(i) We are the first to present a subset of the FactBank dataset containing nested source and target factuality. This allows us to define two related tasks with associated datasets, source-and-target factuality and author-only factuality. We create a database of the complex FactBank corpus for public release.

(ii) We are the first to present a generative machine learning architecture for the factuality prediction task. We perform multiple experiments with factuality structure and target generated text structure, and offer insights into how to frame the event factuality prediction task as a text generation task.

(iii) We perform multi-task learning to improve on both factuality tasks. We offer a detailed evaluation of what combinations work and why.

(iv) We achieve state-of-the-art results in an end-to-end setting for the FactBank source-and-target and author-only factuality tasks.

We first present the problem we are solving (Section 2). We then present a survey of previous work (Section 3). In Section 4, we present the FactBank database architecture. Section 5 details our generative experimental details and modeling framework. Finally, in Sections 6 and 7 we report experiments on the FactBank source-and-target and author-only tasks, respectively.

Findings of the Association for Computational Linguistics: ACL 2023, pages 701–715
July 9-14, 2023 ©2023 Association for Computational Linguistics

2 Background and Motivation

To understand the notion of factuality, consider the following sentence from the FactBank corpus (we have replaced a pronoun for clarity in this exposition). This sentence reports on three events: a selling event, a saying event, and a doubling event. Note that, in this paper, we are not interested in temporal relations, and the notion of factuality applies independently of whether an event is in the past, happening at utterance time, or in the future.

(1) Michael Wu sold the property to five buyers and said he'd double his money.

We can identify four different factuality claims in this sentence:

1. The author is presenting the selling event as factual, i.e., they are committed to the selling event having happened.

2. The author is presenting the saying event as factual, i.e., they are committed to the saying event having happened.

3. The author is presenting the doubling event as having an unknown factuality.

4. The author is presenting Michael Wu as presenting the doubling event as factual, i.e., according to the author, Michael Wu is committed to the doubling event happening.

The first three are claims from the author's perspective, while the last one is from Wu's perspective. We refer to the bearer of the perspective as the *source*, and the event (or state) that the factuality judgment is about as the *target*. FactBank, following MPQA (Wiebe et al., 2005a; Deng and Wiebe, 2015), represents the source of a factuality judgment as an ordered list of sources, since the sentence does not directly tell us about Michael Wu's factuality judgment, but rather the author's claim about Michael Wu's factuality judgment. In this paper, we do not address the explicit reconstruction of such attribution chains.

In the above example, we have seen two factuality values: certain factual, and unknown. We can identify additional values by allowing for non-certain factuality (something may have happened),[1] and by incorporating polarity (something has happened or has not happened). This gives us a set of five possible values for factuality, as shown in Table 1. Thus, we can represent each factuality judgment as a triple consisting of source, target, and factuality value. We represent the source and target by the head words of the corresponding syntactic spans. If the source is the author, we use a special token *AUTHOR*. Our example in (1) can then be represented as follows:

CT-	PR-	UU	PR+	CT+
false	possibly false	unknown	possibly true	true

Table 1: Factuality value mappings for the FactBank corpus

(2)

source	target	facuality value
AUTHOR	sold	CT+
AUTHOR	said	CT+
AUTHOR	double	UU
Wu	double	UU

In NLP, there is a distinct task of determining whether a statement is true or not (fact checking). Unfortunately, this other task is sometimes also called "factuality prediction" (see, for example, (Baly et al., 2018)). The difference is that we are interested in how the author *presents* the event, not ground truth. So despite the same or similar name, there are two different tasks and we only deal with the presentation task, not the ground truth task.

3 Related Work

Author-Only Factuality Corpora All event-factuality corpora focus on the presentation of factuality according to the author of the text, with the exception of FactBank, which also annotates the factuality of the mentioned sources besides the author. These corpora include 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 mainly differ as to what is defined as an annotatable event, the genre of the text, the type of annotators, and the annotation scale. These corpora were unified under a continuous annotation scale in

[1] FactBank divided this category into the probable and the possible, but this leads to data fragmentation, and it can also be hard for humans to distinguish these two cases.

the range [-3, 3] by Stanovsky et al. (2017) (though the author-only factuality value in FactBank was misinterpreted, see (Murzaku et al., 2022) for details).

FactBank The main focus of this paper is the FactBank corpus, which annotates all events introduced in a corpus of exclusively newswire text. The FactBank corpus not only annotates the factuality presented by the author of a text towards an event, but also the factuality of events according to their presentation by sources mentioned inside of the text. Saurí and Pustejovsky (2012) were the first to investigate and perform experiments on the source and target annotations in FactBank. However, we cannot perform an apples-to-apples comparison, as their system neither recognizes events nor identifies sources mentioned in the text. Rather, in their evaluation, this information was created from manual annotation, fed to the system, and then tested on the whole FactBank corpus.

We choose to focus on FactBank because of its expert-level annotations and its detailed source and target annotations. Because of the complexity of the FactBank corpus, we build a robust and efficient database representation of FactBank, which includes all sources including the author, the targets of the factuality attributions, and their respective relations.

Machine Learning Architectures All previous approaches on the event-factuality prediction task use author-only corpora and predict factuality according to the author of the text. Early approaches to the event factuality prediction task used rule-based systems or lexical and dependency tree based features (Nairn et al., 2006; Lotan et al., 2013). Expanding on these rule-based approaches, other work on the event factuality prediction task used SVMs alongside these dependency tree and lexical based features (Diab et al., 2009; Prabhakaran et al., 2010; Lee et al., 2015; Stanovsky et al., 2017). Early neural work includes LSTMs with multi-task or single-task approaches (Rudinger et al., 2018) or using BERT representations alongside a graph convolutional neural network (Pouran Ben Veyseh et al., 2019). Jiang and de Marneffe (2021) expand on these previous works by using other event factuality corpora in multiple training paradigms while also introducing a simpler architecture. These previous neural approaches evaluate on Pearson correlation and mean absolute error (MAE). In previous work, we provide the first end-to-end evaluation using F-measure of the author-only event factuality prediction task (Murzaku et al., 2022).

Our work differs from the previous work in two major ways: first, we are the first to provide a novel and end-to-end generative approach for the event factuality prediction tasks (both author-only and source-and-target). Furthermore, besides our own previous work (Murzaku et al., 2022), all previous works assumed gold event heads. Our system is by default end-to-end, making it usable in real world applications. Second, we perform experiments on the nested sources and target event's factuality, while other works only focused on the presentation of factuality according to the author.

ABSA and ORL Two tasks close in formulation to our task and from which we adopt ideas and insights are the aspect-based sentiment analysis (ABSA) task and the opinion role labelling task (ORL). Peng et al. (2020) create the aspect sentiment triplet extraction task to predict triplets consisting of aspects, opinions, and sentiment polarity. Zhang et al. (2021) are the first to use a generative approach for ABSA fine-tuning on T5. Expanding on this, Gao et al. (2022) achieve state-of-the-art results on all ABSA corpora using a multi-task learning approach through task-specific prompts. The ORL task aims to discover opinions, the sources of opinions, and the associated targets of opinions using the MPQA 2.0 corpus (Wiebe et al., 2005b). Xia et al. (2021) build an end-to-end system creating span representations and using a multi-task learning framework. They achieve state-of-the-art results in the end-to-end setting on the exact match F1 metric.

4 FactBank Database

We present a generalized database structure for capturing cognitive states expressed in language. The goal is to unify multiple annotated corpora in one format, and to make it simple for users to extract the information they need in various formats. In this paper, we describe only how we use it to hold the annotations of event factuality corpora, and of FactBankin particular, whether in the author-only perspective or source-and-target perspective. However, given the diversity of corpora, with each corpus having its own focus, annotation rules, and annotation styles, our database structure is sufficiently broad and abstract to accommodate various corpora equally well and yet to preserve the rich-

ness of information that each corpus offers, so as to facilitate combining corpora in future experiments with as little data loss as possible. Our goal of preserving the distinct details of individual corpora serves as a step in the direction of bringing human knowledge to bear upon otherwise black-box machine learning techniques.

As an example, consider the FactBank and LU (Diab et al., 2009) corpora. The LU data was published as GATE-formatted XML files with annotation targets and annotations given in XML elements, whereas FactBank was published as a set of text files, each of which represents a relation in what amounts to a relational database. From both of these data sources, we may want to construct, for each training and testing example, a set of triples (*sentence*, *target-marked-elements*, *label*), where *target-marked-elements* are the tokens of the sentence that describe the target of the factuality judgment by the author, and to which the *label* refers. If we used the original FactBank release and created a database from it, eliciting triples satisfactory for machine learning wouldl require a complex query with many joins and filters. This is because the structure of the FactBank (implicit) database is oriented toward event-time relations rather than factuality labeling. Accordingly, we designed a new database structure more amenable to queries to support machine learning and developed code to translate corpora including FactBank into this database model.

Database Structure To build the unified database, we needed a stable, fast, and lightweight tool. Python's extensive library support for SQLite database interactions fit those requirements. The unified database's schema is composed of four tables: **sentences**, **mentions**, **sources**, and **attitudes**. We provide a graphic of the database schema in Appendix C.

The **sentences** table stores each sentence and any relevant identifying metadata. Thus far, we have not encountered any corpora with suprasentential information encoded as labels. In principle, however, this table can be refactored to accommodate possible future suprasentential information.

Elements within each sentence marked for labeling are stored in the **mentions** table, with an entry being composed of the surface text of the element, which may be one or more tokens, and its character offset within the sentence. Each sentence may contain more than one marked element.

The **sources** table represents not only sources but their possible nested relations within sentences. These "according-to" relations form a list, as in ***Mary*** *said that* ***John*** *said that* ***Jane*** *was coming to dinner*. Here, the embedded source for the *coming* event is (Author → Mary → John). These "according-to" relations may form a tree, as in ***Mary*** *said that* ***John*** *said that* ***Jane*** *was coming to dinner, but* ***Bob*** *said that she was not*. Here, the embedded source for the *coming* event is (Author → Mary → John). The author may have more than one child source, as in ***Mary*** *said that* ***John*** *was* ***coming*** *to dinner, but* ***Bob*** *said that* ***John*** *was staying home*. Here, we have (Author → Mary) as source for the *coming* event, and (Author → Bob) as source for the *staying* event.

Each sentence may have more than one source, but each source has at most one mention. The implied author has no mention, and a named source mentioned repeatedly is listed once for each mention, since we do not apply anaphora resolution.

Finally, the **attitudes** table aggregates a sentence, its marked elements, and the factuality or sentiment label; the table accommodates both labels but could be refactored to support further label types. Each source may have a distinct attitude toward each of several targets, and each target may have more than one source with its own attitude toward that target. Thus, each source-target pair drawn from mentions has a single listing in attitudes.

Using event factuality corpora annotated on source-and-target factuality is inherently complex and requires structure induction, source linking, and complex database-like operations. Our database structure is an initial step to address the complexity of corpora while also making easy-to-use software for corpus projection and conversion. Our database for FactBank is available at `https://github.com/t-oz/FactBankUniDB`.

5 Task Definitions and Machine Learning Approach

5.1 Task Definitions and Data

Source and Target Factuality (STF) We define the source-and-target factuality task conceptually as the task to generate all *(source, target, factuality label)* triplets for a given input sentence such that the source is *not* the author, the factuality label belongs to a categorial scale, and the source views the target with the given factuality label.

	Triplet P	Triplet R	Triplet F1	Source	Target	S+T
FBST NoN	$0.499_{\pm 0.020}$	$0.448_{\pm 0.023}$	$0.472_{\pm 0.019}$	$0.826_{\pm 0.008}$	$0.701_{\pm 0.002}$	$0.567_{\pm 0.001}$
FBST-AV NoN	$0.516_{\pm 0.033}$	$0.517_{\pm 0.019}$	$0.517_{\pm 0.023}$	$0.879_{\pm 0.004}$	$0.704_{\pm 0.021}$	$0.610_{\pm 0.020}$
FBST N	$0.542_{\pm 0.013}$	$0.486_{\pm 0.027}$	$0.512_{\pm 0.019}$	$0.865_{\pm 0.001}$	$0.715_{\pm 0.009}$	$0.596_{\pm 0.008}$
FBST-AV N	$0.535_{\pm 0.031}$	$0.535_{\pm 0.016}$	$0.535_{\pm 0.021}$	$0.894_{\pm 0.004}$	$0.724_{\pm 0.018}$	$0.620_{\pm 0.018}$

Table 2: Results on triplet generation evaluated on triplet precision, recall, exact match F1, source F1, target F1, and source and target F1 for the FactBank source and target projection (FBST). NoN denotes no normalization, N denotes normalization, and AV denotes attribute-value structure. A shaded cell indicates the best performing combination; light means only a slight improvement.

Projection	Name	Train	Dev	Test
Source-&-Target	FBST	2.5K	767	392
Author-only	FBAO	6.8K	1.9K	1K

Table 3: Information on data set sizes

Author-Only Factuality (AOF) We define the author-only factuality task conceptually as the task to generate all *(event, factuality label)* pairs for a given input sentence such that the factuality label belongs to a categorial scale, and the author views the target with the given factuality label.

For each task, we have created a separate disjoint projection from the full FactBank database. We provide information about these projections in Table 3.

5.2 Representation of Factuality

Previous work represented factuality on a continuous [-3, 3] scale or directly used the categorial factuality labels used in FactBank. We convert the categorial and numerical representation of FactBank to words. We use the word values shown in Table 1 for all experiments containing factuality values, as using the words leads to better task-specific embeddings therefore leading to better performance (on average 5% for our baseline FactBank source and target experiments).

5.3 Input/Output Formats

We define our input x as the raw text and prepend a task prefix p depending on the task of choice. We use a distinct task prefix for each task so that the backbone language model can distinguish between different tasks. For each sub-task that we perform, we define separate target output formats.

Tuple Representation We represent the target as tuples. We use example (1) above to show how this data is represented. For the STF task, the output is a list of triplets:

Input: **source target factuality:** Michael Wu sold the property to five buyers and said he'd double his money.
Output: (Wu, double, true)

For the AOF task, the output is a list of pairs:

Input: **author only factuality:** Michael Wu sold the property to five buyers and said he'd double his money.
Output: (sold, true); (said, true); (double, unknown)

Attribute-Value Representation (AV) As an alternative, we structure our target text in an attribute-value pair format. For the STF task, we get:

Input: **source target factuality:** Michael Wu sold the property to five buyers and said he'd double his money.
Output: (source = Wu, target = double, true)

For the AOF task, we get:

Input: **author only factuality:** Michael Wu sold the property to five buyers and said he'd double his money.
Output: (target = sold, true); (target = said, true); (target = double, unknown)

Inline Representation (Anno) We also represent the AOF task as in-line annotations in the target text representation, since we can anchor the factuality on the target head word. We follow the same annotation format style as Zhang et al. (2021), as the authors found that this text generation target performs well for tuple data representations. We repeat the example from above in this format:

Input: **author only factuality:** He sold the property to five buyers and said he'd double his money.
Output: Michael Wu [sold | true] the property to five buyers and [said | true] he'd [double | unknown] his money.

Note that this in-line annotation format does not work for the STF task, because it relates two dis-

tinct sentence elements to a factuality value.

5.4 Model

5.4.1 Flan-T5

For all experiments, we use the encoder-decoder pre-trained Flan-T5 model (Chung et al., 2022). The Flan-T5 model yields significant improvements on many tasks over the T5 model (Raffel et al., 2020) by adopting an instruction fine-tuning methodology. By formulating the STF and AOF tasks as a text generation task, we can create end-to-end models without a task-specific architecture design.

5.4.2 Multi-task Learning

Models like T5 and Flan-T5 are multi-task in nature by the pre-training objectives. In the pre-training of T5 (Raffel et al., 2020), T5 was trained with a mixture of tasks separated by task specific prefixes. We perform multi-task learning experiments by prepending task specific prefixes for each task as mentioned in Section 5.1. Furthermore, we also perform proportional mixing to sample in proportion to the dataset size.

6 Experiments: Source and Target

In this section, we perform experiments on the STF task. We evaluate exclusively on FBST. Our goal is to achieve the best results on this projection of the corpus.

6.1 Experimental Setup

Datasets and Target Structure We first offer baselines on the FactBank source and target projection (FBST henceforth). We then perform experiments on the target output structure to determine how much influence this has on results. Finally, we perform multi-task learning experiments with the author-only projection of FactBank, CB (De Marneffe et al., 2019), MPQA (Wiebe et al., 2005b), and UW (Lee et al., 2015). All experiments are performed using the STF paradigm defined in Section 5.1, where our task is to generate lists of triplets of format *(source, target, factuality label)*.

Evaluation Our main method of evaluation is the exact match F1 metric. With this metric, a prediction is only correct if all three elements of the triplet match. This metric is directly equivalent to micro-f1 but we refer to it as the exact match F1 in this paper. Furthermore, to assess how much each corpus combination is contributing to the source and target matching of the triplet, we offer F1 scores for the source, target, and the source and target combination.

Experiment Details We use a standard fine-tuning approach on Flan-T5. We fine-tune our models for at most 10 epochs with a learning rate of 3e-4, with early stopping being used if the triplet-F1 did not increase on the dev set. All experiments are averaged over three runs using fixed seeds (7, 21, and 42). We also report the standard deviation over three runs. We leave more experimental details to Appendix B.

Text Normalization Following insights and methodology from Zhang et al. (2021), we apply their text normalization strategy on our experiments (denoted NoN for no normalization, N for normalized). Zhang et al. (2021) found that text normalization helps for detecting aspect and opinion phrases in *(aspect, opinion, sentiment)* triplets mainly through producing the correct morphology of a word and through addressing orthographic alternatives to words. Their method finds the replacement word from a corresponding vocabulary set using the Levenshtein distance. We note that in our experiments, most of the improvements that normalization yielded were due to correcting morphological errors (e.g. gold is *houses*, model predicts *house*) or capitalization errors (gold is *Mary*, model predicts *mary*).

6.2 Results: Baseline and Target Output Restructuring

Baselines Table 2 shows our baseline results for the FactBank source and target projection. We notice some particular trends in this task and offer insights. First, we see that normalization helps. For our baseline FBST NoN experiment, we report a triplet F1 of 0.472, whereas after normalization, the triplet F1 increases to 0.512. Intuitively, normalization most helps for sources. One of the main benefits of normalization is producing the correct morphology and orthography. We find that FactBank sources are often nouns or proper nouns and normalization ensures the correct orthography. Furthermore, we see that source outperforms target in all cases and that labelling the correct source and target pairs is not a trivial task. These results are similar to Xia et al. (2021) who worked on the MPQA corpus, which annotates opinions (i.e., text passages indicating opinions), sources of opinions, and the targets of these events. The authors found

Combo	Triplet P	Triplet R	Triplet F1	Source	Target	S+T
Baseline: FBST	0.535±0.031	0.535±0.016	0.535±0.021	0.894±0.004	0.724±0.018	0.620±0.018
FBST, CB	0.562±0.017	0.536±0.017	0.549±0.024	0.907±0.013	0.729±0.002	0.633±0.008
FBST, MPQA	0.497±0.009	0.485±0.009	0.491±0.030	0.903±0.020	0.715±0.007	0.615±0.023
FBST, UW	0.585±0.013	0.526±0.013	0.553±0.010	0.882±0.010	0.725±0.013	0.631±0.002
FBST, FBAO	0.683±0.025	0.655±0.025	0.669±0.032	0.890±0.029	0.854±0.009	0.746±0.014
FBST, FBAO*	0.710±0.030	0.661±0.030	0.684±0.030	0.893±0.005	0.837±0.012	0.753±0.010

Table 4: Results on triplet precision, recall, exact match F1, source F1, target F1, and the source and target F1 for the MTL experiments on generating factuality triplets for the FactBank source and target projection (FBST). A shaded cell indicates state-of-the-art; light means only a slight improvement.

	Macro-F1	CT+	PR+	UU	PR-	CT-
Murzaku et al. (2022)	0.680	0.767	0.714	0.735	0.667	0.519
FBAO	0.604±0.094	0.891±0.016	0.317±0.152	0.754±0.016	0.389±0.347	0.667±0.039
FBAO-Anno	0.632±0.065	0.791±0.010	0.436±0.139	0.774±0.005	0.389±0.347	0.769±0.059
FBAO-Pol	0.667±0.023	0.907±0.030	0.334±0.059	0.792±0.020	0.667±0.000	0.695±0.065
FBAO-Anno-Pol	0.690±0.008	0.792±0.003	0.246±0.060	0.751±0.013	1.000±0.000	0.685±0.041
FBAO*, FBST	0.694±0.029	0.939±0.015	0.312±0.036	0.809±0.008	0.778±0.192	0.675±0.061

Table 5: Results on FactBank author-only (FBAO) compared to the end-to-end SOTA held by Murzaku et al. (2022). We show results for the in-line annotation style (FBAO-Anno) and the result modelling our task alongside polarity (FBAO-Pol). A shaded cell indicates a new SOTA; light means only a slight improvement.

that matching MPQA sources to opinions is far easier than matching MPQA targets to opinions.

Attribute-Value (AV) Addition In Table 2, we also report results on experiments where we use the attribute-value (AV) format for the output. This formatting especially helps with disambiguation of the source, targets, and factuality, providing our generative framework deeper contextual understanding and cues for triplet generation. We find that this output format produces large increases in all measures, namely the triplet F1, source F1, and source and target F1. Once again, we see that normalization helps, achieving our highest baseline triplet F1 of 0.535. Because of the success of this target format restructuring (AV) and normalization (N), we perform the remaining experiments in this paper using the AV output format and the normalization step.

6.3 Results: Multi-task learning experiments

We perform multi-task learning (MTL) experiments using author-only factuality corpora, opinion role labelling corpora, and the combinations of all of them. Following our approach described in Section 5.4.1, we prepend task specific prefixes for our tasks, such as **author only factuality:** or **opinion role label:** . We mirror the format of our FactBank source and target examples for our MTL experiments. For example, when we add in the author-only factuality data, we structure our targets as *(target = event, factuality label)*, mirroring the format of our source and target data. Similarly, for other corpora such as MPQA which only contain source and target information without any factuality labels, we structure our data as *(source = opinion source, target = opinion target)*. We aim to tackle the following with our MTL experiments: first, we aim to improve target identification. Our FBST-only system performs worse on identifying targets than sources. To address this, we combine with author-only event factuality corpora, namely FactBank (denoted FBAO), and CB and UW, which both annotate events in a similar structure and genre as FactBank. Second, we aim to improve source and target linking, as the FBST-only system cannot perform well on this task. We attempt to address this using the Xia et al. (2021) projection of the MPQA corpus which annotates opinion sources and opinion targets. We also attempt an experiment with a direct mirroring of the source and target representation when using the FactBank author-only data (we denote this representation as FBAO*). Here, we explicitly state the author of the text as a source, structuring our target

text to be generated as *(source = AUTHOR, target = event, factuality label)*.

Results for our MTL experiments are shown in Table 4. We see that all corpus combinations besides MPQA help for the triplet F1 metric. Most notably, we find that adding the FactBank author-only data (FBAO) and in particular, the triplet FactBank author-only projection (FBAO*) helps the most, especially for the target and source+target F1. We note though that the triplet F1 results for FBST with FBAO and FBAO* both have rather large standard deviations, so the difference may not be significant. Adding other author-only factuality corpora such as UW and CB help, but not as much as FactBank. We see that CB does not boost performance much on FactBank, and UW actually helps more for the triplet F1 metric. This may be because we are performing a separate task and using a different machine learning paradigm. MPQA does not help for any metric besides the source metric. Opinion role labelling is a separate task and appears to be incompatible with the source and target factuality task. However, we note that MPQA also annotates targets differently from FactBank, which explains why the MTL approach did not help in this case.

7 Experiments: Author Only

In this section, we perform experiments on the AOF task. We evaluate exclusively on FBAO, performing our experiments with the same model and training paradigm. We use three styles of target representation mentioned in Section 5: one style where we extract event words and their associated factuality values as tuples, an in-line annotation style used by Zhang et al. (2021), and finally a MTL triplet generation task with the source and target projection of FactBank where we generate triplets of format *(source = AUTHOR, target = target event, factuality label)*. Furthermore, we also factor polarity in our experiments. Murzaku et al. (2022) found that separately predicting polarity and factuality for the event factuality task can lead to error reductions since polarity is often expressed independently of the degree of factuality. We treat the addition of polarity as a triplet generation task generating triplets of format *(target = target event, factuality label, polarity)*. We reduce the factuality label to the strength of factuality (true, possibly true, unknown), with the polarity being one of (negative, unknown, positive).

7.1 End-to-End Author-Only Factuality

We follow the end-to-end evaluation setup on FactBank as we did in (Murzaku et al., 2022), evaluating on per-label F1 and macro-F1. Because our system is end-to-end, we cannot evaluate on Pearson correlation or MAE like some previous event factuality papers that assumed gold heads. For an apples-to-apples comparison, we use the same label mappings as Murzaku et al. (2022). We average over three runs and also report standard deviation which the previous authors did not report.

Table 5 shows results for our experiments on FactBank author-only (FBAO), FBAO with an in-line annotation target format (FBAO-Anno), FBAO as a triplet generation task that includes polarity (Pol), and FBAO finally a MTL triplet generation task with the source and target projection of FactBank, tested on FBAO (FBAO*, FBST). We note the very high standard deviations in the PR+ and PR- measurements; these labels are rare even after collapsing them to the same class, especially in the test set, which explain the extreme standard deviation fluctuations. Our baseline system (FBAO) yields a noticeable increase in the CT+, UU, and CT- labels compared to the baseline, but performs worse on the PR+ and PR- labels. The in-line annotation text generation task performs better on macro-F1 than the baseline tuple generation task, with a notable increase in CT-. Factoring polarity helps as well: for both configurations, factoring polarity leads to an increase and achieves new a SOTA for the PR- label in our FBAO-Anno-Pol setup. Our best performing result is our multi-task learning on FBAO and FBST, where we modify FBAO to include the author as a source in its triplet representation. We achieve new SOTA on macro-f1, a large increase and SOTA on the CT+ label, and SOTA on UU.

7.2 FBAO: Exact Match Evaluation

To be able to compare performance on the STF and AOF tasks, we evaluate using the same metric as Section 6, specifically using tuple/triplet exact match precision, recall, F1, and target F1. This evaluation corresponds to a micro-F1, as it does not depend on the factuality value. In this evaluation, we do not consider source F1 or source and target F1 because the source is the author of the text. We aim to quantify how well our generative system performs at generating author-only structures, and therefore evaluate using an exact match evaluation.

	P	R	EM F1	Target F1
FBAO	0.858 ±0.004	0.874 ±0.012	0.866 ±0.007	0.865 ±0.004
FBAO-Anno	0.789 ±0.004	0.750 ±0.013	0.769 ±0.009	0.845 ±0.001
FBAO-Pol	0.878 ±0.016	0.892 ±0.021	0.884 ±0.018	0.884 ±0.001
FBAO-Anno Pol	0.786 ±0.006	0.750 ±0.008	0.767 ±0.005	0.849 ±0.002
FBAO*, FBST	0.895 ±0.009	0.898 ±0.009	0.897 ±0.008	0.889 ±0.003

Table 6: Results on FactBank author-only (FBAO) using a precision, recall, tuple exact match F1, triplet exact match F1 for the FBAO* and FBST combo, and target F1. A shaded cell indicates a new SOTA; light means only a slight improvement.

We are the first to report results on FactBank using an exact match evaluation.

Table 6 shows results for our exact match evaluation on FBAO. We see two clear trends: first, the in-line annotation generation task does not perform as well in our exact match evaluation compared to our tuple/triplet generation task. This makes sense given that the Anno option performs markedly worse on the most common factuality value, CT+, which in the macro-average is compensated by better performances for other values, but in the exact-match evaluation lowers its overall performance. Our best results are produced by our MTL setup with FBAO and FBST(FBAO*, FBST). Similar to our source and target results in Table 4, we see that the AOF task benefits from the FBST data in a MTL setup performing the best once again. We also see, as expected, that the AOF task is easier than the STF task, with a result margin of 13.3% absolute, since fewer details need to be predicted, and since more data is available.

8 Conclusion

We provide a new generative framework for the event factuality prediction task using Flan-T5 and focusing on output format, individual task prefixes, and multi-task learning. To tackle the complexity of the FactBank corpus, we create a database representation that simplifies extracting sources, targets, and factuality values for all projections of FactBank, which we will publicly release. Our source-and-target experiments show that careful output formatting can yield improvements (Table 2) and careful attention to multi-task learning mixtures can help (Table 4). We evaluate the author-only event factuality task using both macro-average (Table 5) and exact-match evaluation metrics (Table 6), with as expected different results. We achieve new state-of-the-art results on both source-and-target (because no prior results) and author-only (beating existing results) end-to-end factuality prediction.

Acknowledgements

We thank the anonymous reviewers for their helpful insights and comments. 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. Rambow gratefully acknowledges support from the Institute for Advanced Computational Science at Stony Brook University.

Limitations

While we achieved preliminary results and created a preliminary projection of the FactBank source and target corpus, we do not capture the full source and target nesting in our machine learning experiments. We repeat the example from Section 4: ***Mary*** *said that* ***John*** *said that* ***Jane*** *was coming to dinner, but* ***Bob*** *said that she was not*. The embedded sources for the *coming* event are (Author → Mary → John), which translates to "according to the author according to Mary according to John, did the coming event happen?" In our experiments and machine learning architecture, we focus on the last nested source, or John in this example. In future work, we aim to link together all sources and their embedded nesting structures.

We note that all experiments in this paper were performed using the Flan-T5-base model. In future work on this task, we will explore different generative models such as GPT-3 or BART, which may yield stronger performing systems or more interesting results. We are especially curious about framing this task using GPT-3, especially performing tasks on few-shot or in-context learning.

Finally, we note that these experiments do not account for potential biases prevalent in fine-tuning large language models. We hypothesize that for some sources in text (i.e. power figures, authorities, or specific names), there may be biases towards certain labels. We will investigate these biases in future work, as an event factuality prediction system with inherent bias can have real world implications.

Ethics Statement

As mentioned in the limitations section, we note that these experiments do not account for potential biases prevalent in fine-tuning large language models. In a real world deployment of our model, we hypothesize that there could be a potential mislabelling of factuality values depending on bias towards sources of utterances. For example, if a power figure states an event, will the event label be more biased towards being factual just because of the source of the statement? We will investigate these questions and issues in future work.

We also note that our paper is foundational research and we are not tied to any direct applications.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

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.

William Falcon et al. 2019. Pytorch lightning. *GitHub. Note: https://github. com/PyTorchLightning/pytorch-lightning*, 3(6).

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

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).

John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. Re-examining FactBank: Predicting the author's presentation of factuality. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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)*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment

analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8600–8607.

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

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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(3):227–268.

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.

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, Vancouver, Canada. Association for Computational Linguistics.

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

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005a. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005b. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Qingrong Xia, Bo Zhang, Rui Wang, Zhenghua Li, Yue Zhang, Fei Huang, Luo Si, and Min Zhang. 2021. A unified span-based approach for opinion mining with syntactic constituents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1795–1804, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

A Distribution of Data Set and Database

We intend to distribute the split of the source and target FactBank dataset. We have included the dataset in this submissions for reviewers to inspect, but cannot distribute it due to copyright reasons. Instead, we will release a Python script alongside our SQLite database implementation which will produce the files submitted with this paper with the original FactBank corpus as an input. The FactBank corpus can be obtained by researchers from the Linguistic Data Consortium, catalog number LDC2009T23.

Our dataset split is detailed in Table 3. We split our corpus using the methods as Murzaku et al. (2022), which also includes splitting by article.

B Details on Experiments

We use a standard fine-tuning approach on the Flan-T5-base model with 247,000,000 parameters. For computing, we used our employer's GPU cluster and performed experiments on a Tesla V100-SXM2 GPU. Compute jobs typically ranged from 10 minutes for small single corpus combinations, to 30 minutes for larger multi-task learning corpus combinations. We did not do any hyperparameter search or hyperparameter tuning.

We fine-tuned our models for at most 10 epochs with a learning rate of 3e-4, with early stopping being used if the triplet-F1 did not increase or if the factuality macro-F1 did not increase. All metrics for experiments were averaged over three runs using fixed seeds (7, 21, and 42). We report the average over three runs and the standard deviation over three runs.

For prediction normalization on our fixed experiments setting, we use the editdistance Python package. We provide scripts for our prediction normalization and full evaluation and will be made publicly available.

To fine-tune our models and run experiments, we used PyTorch lightning Falcon et al. (2019) and the transformers library provided by HuggingFace Wolf et al. (2019). All code for fine-tuning, modelling, and preprocessing will be made available.

C Database Structure

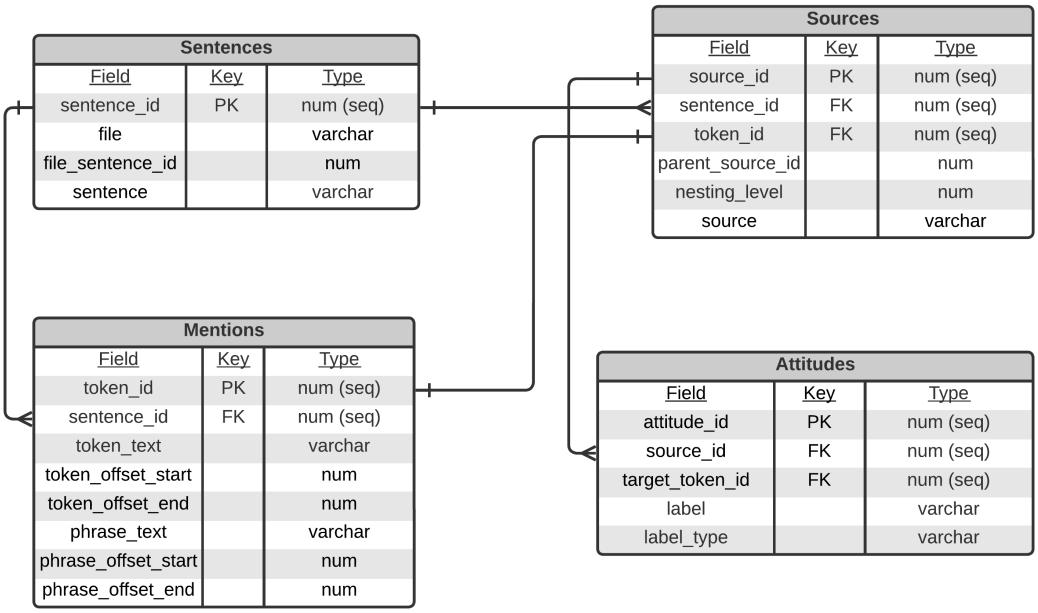

Figure 1: Entity-Relation Diagram of the FactBank Database. Note that the one-to-one notation between mentions and sources only applies to source mentions, not target mentions, which are one-to-many.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discussed limitations in the Limitations section.

☑ A2. Did you discuss any potential risks of your work?
We discussed risk in the Ethics Statement section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We summarize the paper's main claims and contributions in the abstract and section 1 (Introduction).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Sections 1, 2, and 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
All data originated from the FactBank corpus.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 1 and 2.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We show relevant corpus specific statistics in table 3 and mention it in section 5.

C ☑ Did you run computational experiments?

Sections 6 and 7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We provide the average over 3 runs and standard deviations in all tables in Sections 6 and 7.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We used PyTorch lightning for training our models. We describe the setup in section 5 and the experimental details in Appendix B.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.