Right for the Right Reason: Evidence Extraction for Trustworthy Tabular Reasoning

Vivek Gupta[1]*, **Shuo Zhang**[2], **Alakananda Vempala**[2],
Yujie He[2], **Temma Choji**[2], **Vivek Srikumar**[1]
[1]University of Utah, [2]Bloomberg,
{vgupta, svivek}@cs.utah.edu, {szhang611, avempala, yhe247, tchoji}@bloomberg.net

Abstract

When pre-trained contextualized embedding-based models developed for unstructured data are adapted for structured tabular data, they perform admirably. However, recent probing studies show that these models use *spurious* correlations, and often predict inference labels by focusing on false evidence or ignoring it altogether. To study this issue, we introduce the task of *Trustworthy Tabular Reasoning*, where a model needs to extract evidence to be used for reasoning, in addition to predicting the label. As a case study, we propose a *two-stage sequential prediction* approach, which includes an *evidence extraction* and an *inference* stage. First, we crowdsource evidence row labels and develop several unsupervised and supervised evidence extraction strategies for INFOTABS, a tabular NLI benchmark. Our evidence extraction strategy outperforms earlier baselines. On the downstream tabular inference task, using only the automatically extracted evidence as the premise, our approach outperforms prior benchmarks.

Breakfast in America		Relevant
Released[4]	29 March 1979[4]	H3
Recorded[3,4]	May-December 1978[3,4]	H2, H3
Studio	The Village Recorder in Los Angeles[3]	
Genre	Pop, Art Rock, Soft Rock	
Length[2]	46:06[2]	H1
Label	A&M	
Producer[1]	Peter Henderson, Supertramp[1]	H1

H1: Supertramp produced[1] an album that was less than an hour long[2].
H2: Most of Breakfast in America was recorded[3] in the last month of 1978[3].
H3: Breakfast in America was released[4] the same month recording ended [4].

Figure 1: A semi-structured premise (the table 'Breakfast in America') example from (Gupta et al., 2020). Hypotheses H1 are entailed by it, H2 is neither entailed nor contradictory, and H3 is a contradiction. The Relevant column shows the hypotheses that use the corresponding row. The colored text (and superscripts) in the table and hypothesis highlights relevance token level alignment.

1 Introduction

Reasoning on tabular or semi-structured knowledge is a fundamental challenge for today's natural language processing (NLP) systems. Two recently created tabular Natural language Inference (NLI) datasets, TabFact (Chen et al., 2020b) on Wikipedia relational tables and INFOTABS (Gupta et al., 2020) on Wikipedia Infoboxes help study the question of inferential reasoning over semi-structured tables. Today's state-of-the-art for NLI over unstructured text uses contextualized embeddings (e.g., Devlin et al., 2019; Liu et al., 2019b). When adapted for tabular NLI by flattening tables into synthetic sentences using heuristics, these models achieve remarkable performance on the datasets.

However, a recent study (Gupta et al., 2021) demonstrates that these models fail to reason properly on the semi-structured inputs in many cases. For example, they can ignore relevant rows, and (a) focus on the irrelevant rows (Neeraja et al., 2021), (b) use only the hypothesis sentence (Poliak et al., 2018; Gururangan et al., 2018), or (c) knowledge acquired during pre-training (Jain et al., 2021; Gupta et al., 2021) . In essence, they use spurious correlations between irrelevant rows, the hypothesis, and the inference label to predict labels.

This paper argues that existing NLI systems optimized solely for label prediction cannot be trusted. It is not sufficient for a model to be merely *Right* but also *Right for the Right Reasons*. In particular, at least identifying the relevant elements of inputs as the *'Right Reasons'* is essential for trustworthy reasoning[1]. We address this issue by introducing

*Work done during an internship at Bloomberg

[1] We argue that a reasoning system can be deemed trustworthy only if it exposes how its decisions are made, thus admitting verification of the reasons for its decisions.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3268 - 3283

the task of *Trustworthy Tabular Inference*, where the goal is to extract relevant rows as evidence and predict inference labels.

To illustrate this task, consider an example from the INFOTABS dataset in Figure 1, which shows a premise table and three hypotheses. The figure also marks the rows needed to make decisions about each hypothesis, and also indicates the relevant tokens for each hypothesis. For trustworthy tabular reasoning, in addition to predicting the label ENTAIL for *H1*, CONTRADICT for *H2* and NEUTRAL for *H3*, the model should also identify the evidence rows—namely, the rows *Producer* and *Length* for hypothesis *H1*, *Recorded* for hypothesis *H2*, *Released* and *Recorded* for hypothesis *H3*.

As a first step, we propose a two-stage sequential prediction approach for the task, comprising of an evidence extraction stage, followed by an inference stage. In the evidence extraction stage, the model extracts the necessary information needed for the second stage. In the inference stage, the NLI model uses only the extracted evidence as the premise for the label prediction task.

We explore several unsupervised evidence extraction approaches for INFOTABS. Our best unsupervised evidence extraction method outperforms a previously developed baseline by 4.3%, 2.5% and 5.4% absolute score on the three test sets. For supervised evidence extraction, we annotate the INFOTABS training set (17K table-hypothesis pairs with 1740 unique tables) with relevant rows following the methodology of Gupta et al. (2021), and then train a RoBERTa$_{\text{LARGE}}$ classifier. The supervised model improves the evidence extraction performance by 8.7%, 10.8%, and 4.2% absolute scores on the three test sets over the unsupervised approach. Finally, for the full inference task, we demonstrate that our two-stage approach with best extraction, outperforms the earlier baseline by 1.6%, 3.8%, and 4.2% on the three test sets.

In summary, our contributions are as follows[2]:

- We introduce the problem of trustworthy tabular reasoning and study a two-stage prediction approach that first extracts evidence and then predicts the NLI label.
- We investigate a variety of unsupervised evidence extraction techniques. Our unsupervised approach for evidence extraction outperforms the previous methods.
- We enrich the INFOTABS training set with evidence rows, and develop a supervised extractor that has near-human performance.
- We demonstrate that our two-stage technique with best extraction outperforms all the prior benchmarks on the downstream NLI task.

[2] The updated dataset, along with associated code, is available at https://tabevidence.github.io/.

2 Task Formulation

We begin by introducing the task and the datasets we use.

Tabular Inference is a reasoning task that, like conventional NLI (Dagan et al., 2013; Bowman et al., 2015; Williams et al., 2018), asks whether a natural language *hypothesis* can be inferred from a tabular *premise*. Concretely, given a premise table T with m rows $\{r_1, r_2, \ldots, r_m\}$, and a hypothesis sentence H, the task maps them to ENTAIL (E), CONTRADICT (C) or NEUTRAL (N). We can denote the mapping as

$$f(\text{T}, \text{H}) \rightarrow \text{y} \tag{1}$$

where, $\text{y} \in \{\text{E}, \text{N}, \text{C}\}$. For example, for the tabular premise in Figure 1, the model should predict E, C, and N for the hypotheses *H1*, *H2*, and *H3*, respectively.

Trustworthy Tabular Inference is a table reasoning problem that seeks not just the NLI label, but also relevant evidence from the input table that supports the label prediction. We use T^R, a *subset* of T, to denote the relevant rows or evidence. Then, the task is defined as follows.

$$f(\text{T}, \text{H}) \rightarrow \{\text{T}^R, \text{y}\} \tag{2}$$

In our example table, this task will also indicate the evidence rows T^R of *Producer* and *Length* for hypothesis *H1*, *Recorded* for hypothesis *H2*, and *Released* and *Recorded* for hypothesis *H3*.

While the notion of evidence is well-defined for the ENTAIL and CONTRADICT labels, the NEUTRAL label requires explanation. To decide on the NEUTRAL label, one must first search for relevant rows (if any), i.e., identify evidence in the premise tables. In fact, this is a causally correct sequential approach. Indeed, INFOTABS has multiple neutral hypotheses that are partly entailed by the table; if any part of a hypothesis contradicts the table, then the inference label should be CONTRADICT. For example, in our example table, the premise table indicates that the album was recorded in 1978, emphasizing the importance of the *Recorded* row for

the hypothesis H2. For NEUTRAL examples, we refer to any such pertinent rows as evidence.

Dataset Details. There are several datasets for tabular NLI: TabFact, INFOTABS, and the SemEval'21 Task 9 (Wang et al., 2021b) and the FEVEROUS'21 shared task (Aly et al., 2021) datasets. We use the INFOTABS data in this work. It contains finer-grained annotation (e.g., TabFact lacks NEUTRAL hypotheses) and more complex reasoning than the others[3].

The dataset consists of 23,738 premise-hypothesis pairs collected via crowdsourcing on Amazon MTurk. The tabular premises are based on 2,540 Wikipedia Infoboxes representing twelve diverse domains, and the hypotheses are short statements paired with NLI labels. All tables contain a *title* followed by two columns (cf. Figure 1); the left columns are *keys* and the right ones are *values*).

In addition to the train and development sets, the data includes multiple test sets, some of which are adversarial: α_1 represents a standard test set that is both topically and lexically similar to the training data; α_2 hypotheses are designed to be lexically adversarial[4]; and α_3 tables are drawn from topics unavailable in the training set. The dev and test set, comprising of 7200 table-hypothesis pairs, were recently extended with crowdsourced evidence rows (Gupta et al., 2021). As one of our contributions, we describe the evidence rows annotation for the training set in the next Section 3.

3 Crowdsource Evidence Extraction

This section describes the process of using Amazon MTurk to annotate evidence rows for the 16,538 premise-hypothesis pairs that make the training set of INFOTABS. We followed the protocol of Gupta et al. (2021): one table and three distinct hypotheses formed a HIT. For each of the hypotheses, five annotators would select the evidence rows. We divide the tasks equally into 110 batches, each batch having 51 HITs each having three examples. To reduce bias induced by a link between the NLI label and row selection, we do not reveal the labels to the annotators. The quality control details are provided in the Appendix §B.

In total, we collected 81,282 annotations from 90 distinct annotators. Overall, twenty five annotators completed over 1000 tasks, corresponding to 87.75 % of the examples, indicating a tail distribution with the annotations. Overall, 16,248 training set table-hypothesis pairs were successfully labeled with the evidence rows[5]. On average, we obtain 89.49% F1-score with equal precision and recall for annotation agreement when compared with majority vote. Furthermore, 85% examples have an F1-score of >80 %, and 62% examples have an F1-score of >90 %. Around 60% examples have either perfect (100%) precision or recall, and 42% have both. Table 1 reports the Fleiss' Kappa score with annotation percentage. The average Kappa score is 0.79 with standard deviation of 0.23[6].

Agreement	Range	Percentage (%)
Poor	< 0	0.27
Slight	0.01 – 0.20	1.61
Fair	0.21 – 0.40	5.69
Moderate	0.41 - 0.60	13.89
Substantial	0.61 - 0.80	22.92
Perfect	0.81 - 1.00	55.61
Overall	mean 0.79	s.t.d. 0.23

Table 1: Examples (%) for each Fleiss' Kappa score bucket.

Choice of Semi-structured Data. The rows of an Infobox table are semantically distinct, though all connected to the title entity. Each row can be considered a separate and uniquely distinct source of information about the title entity. Because of this property, the problem of evidence extraction is well-formed as relevant row selection. The same is not valid for unstructured text, whose units of information may be tokens, phrases, sentences or entire paragraphs, and is typically unavailable (Ribeiro et al., 2020; Goel et al., 2021; Mishra et al., 2021; Yin et al., 2021).

4 Trustworthy Tabular Inference

Trustworthy inference has an intrinsic sequential causal structure: extract evidence first, then predict the inference label using the extracted evidence data, knowledge/common sense, and perhaps formal reasoning (Herzig et al., 2021; Paranjape et al., 2020)[7]. To operationalize this intuition, we chose a two-stage sequential approach which consists of an evidence extraction followed by the NLI classi-

[3] As per Gupta et al. (2020), 33% of examples in INFOTABS involve multiple rows. The dataset covers all the reasoning types present in the Glue (Wang et al., 2018) and SuperGlue (Wang et al., 2019) benchmarks. [4] i.e. minimally perturbing hypothesis to flipped ENTAIL to CONTRADICT label and vice-versa.

[5] We exclude certain example pairings from our training sets since they could not achieve satisfactory agreement after adding more annotators or have label imbalance issues i.e. more the required number of neutrals. [6] We also manually examined hypothesis phrases that signal relevant rows. See Appendix D for details. [7] See more details discussion in §7.

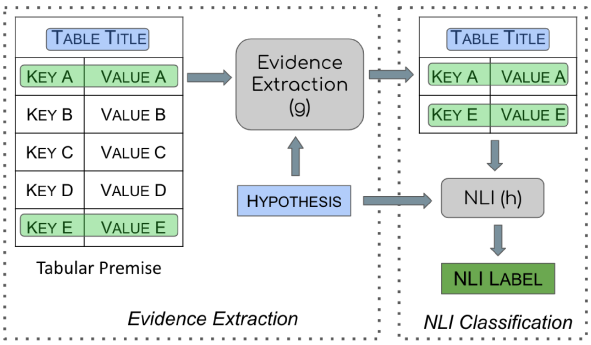

Figure 2: High level flowchart showing our approach for evidence extraction and trustworthy tabular inference.

fication, as shown in Figure 2.

Notation. The function f in Eq. 2 can be rewritten with functions g and h, $f(.) = g(.), h \circ g(.)$, as

$$f(\mathrm{T}, \mathrm{H}) = \{g(\mathrm{T}, \mathrm{H}) \, , \, h\,(g(\mathrm{T}, \mathrm{H}), \mathrm{H})\} \quad (3)$$

Here, g extracts the evidence rows T^R subset of T, and h uses the extracted evidence T^R and the hypothesis H to predict the inference label y, as

$$\begin{aligned} g(\mathrm{T}, \mathrm{H}) &\rightarrow \mathrm{T}^R \\ h(\mathrm{T}^R, \mathrm{H}) &\rightarrow \mathrm{y} \end{aligned} \quad (4)$$

To obtain f, we need to define the functions g and h, and a flexible representation of a semi-structured table T. To represent a table T, we use the **B**etter **P**aragraph **R**epresentation (BPR) heuristic of Neeraja et al. (2021). BPR uses hand-crafted rules based on the table category and entity type's of the row *values* (e.g., boolean and date) to convert each row to a sentence, consisting of table title, key and values. This representation outperforms the original "*para*" representation technique of Gupta et al. (2020).

We explore unsupervised (§4.1) and supervised (§4.2) methods for the evidence row extractor g.

4.1 Unsupervised Evidence Extraction

The unsupervised approaches extract Top-K rows are based on relevance scores, where K is a hyperparameter. We use the cosine similarity between the row and the hypothesis sentence representations to score rows. We study three ways to define relevance described next.

4.1.1 Using Static Embeddings

Inspired by the **D**istracting **R**ow **R**emoval (**DRR**) heuristic of Neeraja et al. (2021), we propose **DRR (Re-Rank + Top-S**$_\tau$**)**, which uses fastText (Joulin et al., 2016; Mikolov et al., 2018) based static embeddings to measure sentence similarity. We employ three modifications to improve DRR.

Re-Rank (δ): We observed that the raw similarity scores (i.e., using only fastText) for some valid evidence rows could be low, despite exact word-level lexical matching with the row's *key* and *values*. We augmented the scores by δ for each exact match to incentivize precise matches.

Sparse Extraction (S): For most instances, the number of relevant rows (K) is much lower than the total number of rows (m); most examples have only one or two relevant rows. We constrained the *sparsity* in the extraction by capping the value of K to S $\ll$ m.

Dynamic Selection (τ): We use a threshold τ to select rows dynamically Top-K$_\tau$ based on the hypothesis, rather than always selecting fixed K rows. We only select rows whose similarity (after Re-Ranking) to the hypothesis sentence representations is greater than a threshold τ. We adopt this strategy because (a) the number of rows in the premise table can vary across examples, and (b) different hypotheses may require a differing number of evidence rows.

4.1.2 Using Word Alignments

This approach consists of two parts (a) aligning rows and hypothesis words, and (b) then computing cosine similarity between the aligned words. Specifically, we use the SimAlign (Jalili Sabet et al., 2020) method for word-level alignment. SimAlign uses static and contextualized embeddings without parallel training data to get word alignments. Among the approaches explored by SimAlign, we use the **Match (mwmf)** method, which uses **m**aximum-**w**eight **m**aximal matching in the bipartite weighted network formed by the word level similarity matrix. Our choice of this approach over the other greedy methods (Itermax and Argmax) is motivated by the fact that it finds the global optimum matching, while the other two do not. After alignment, we normalize the sum of cosine similarities of RoBERTa$_{\text{LARGE}}$ token embeddings[8] to derive the *relevance score*. Furthermore, because all rows use the same title, we assign title matching terms zero weight. This paper refers to this method as SimAlign (Match (mwmf)).

[8] We use the average BPE token embeddings as the word embeddings.

4.1.3 Using Contextualised Embeddings

The approach we saw in §4.1.2 defines row-hypothesis similarity using word alignments. As an alternative, we can directly compute similarities between the contextualised sentence embeddings of rows and the hypothesis. We explore two options here.

Sentence Transformer: We use Sentence-BERT (Reimers and Gurevych, 2019) and its variants (Reimers and Gurevych, 2020; Thakur et al., 2021; Wang et al., 2021a), which use Siamese neural networks (Koch et al., 2015; Chicco, 2021). We explore several pre-trained sentence transformers models[9] for sentence representation. These models differ in (a) the data used for pre-training, (b) the main model type and it size, and (c) the maximum sequence length.

SimCSE: SimCSE (Gao et al., 2021) uses a contrastive learning to train sentence embeddings in both unsupervised and supervised settings. The former is trained to take an input sentence and reconstruct it using standard dropout as noise. The latter uses example pairs from the MNLI dataset (Williams et al., 2018) with entailments serving as positive examples and contradiction serving as hard negatives for contrastive learning.

We give the row sentences directly to SimCSE to get their embeddings. To avoid misleading matches between the hypothesis tokens and those in the premise title, we swap the hypothesis title tokens with a single token title from another randomly selected table of the same category. We then use the cosine similarity between SimCSE sentence embeddings to compute the final relevance score. We again use the sparsity and dynamic selection as earlier. In the study, we refer to this method as SimCSE (Hypo-Title-Swap + Re-rank + Top-K$^{\tau}$).

4.2 Supervised Evidence Extraction

The supervised evidence extraction procedure consists of three aspects: (a) Dataset construction, (b) Label balancing, and (c) Classifier training.

Dataset Construction. We use the annotated relevant row data (§3) to construct a supervised extraction training dataset. Every row in the table, paired with the hypothesis, is associated with a binary label signifying whether the row is relevant or not. As before, we use the sentences from Better Paragraph Representation (BPR) (Neeraja et al., 2021) to represent each row.

Label Balancing. Our annotation, and the perturbation probing analysis of Gupta et al. (2021)[10], show that the number of irrelevant rows can be much larger than the relevant ones for a table-hypothesis pair. Therefore, if we use all irrelevant rows from tables as negative examples, the resulting training set would be imbalanced, with about $6\times$ more irrelevant rows than relevant rows.

We investigate several label balancing strategies by sub-sampling irrelevant rows for training. We explore the following schemes: (a) taking all irrelevant rows from the table without sub-sampling (on average $6\times$ more irrelevant rows) referred to as **Without Sample**($6\times$), (b) randomly sampling unrelated rowsin the same proportion as relevant rows, referred to as **Random Negative**($1\times$), (c) using the unsupervised DRR (Re-Rank + Top-S$_{\tau}$) method to pick the irrelevant rows that are most similar to the hypothesis, in equal proportion as the relevant rows, referred to as **Hard Negative**($1\times$), and (d) same as (c), except picking three times as many irrelevant rows, referred to as **Hard Negative**($3\times$)[11].

Classifier Training. We train a relevant-vs-irrelevant row classifier using RoBERTa$_{\text{LARGE}}$'s two sentence classifier. We use RoBERTa$_{\text{LARGE}}$ because of its superior performance over other models in preliminary experiments, and also the fact that it is also used for the NLI classifier.

4.3 Natural Language Inference

For the downstream NLI task, the function h is a two-sentence classifier whose inputs are T^R (the rows selected by g) and the hypothesis H. We use BPR to represent T^R as we did for the full table T. Since $|\text{T}^R| \ll |\text{T}|$, the extraction benefits larger tables (especially in α_3 set) which exceed the model's token limit.

5 Experimental Evaluation

Our experiments assess the efficacy of evidence extraction (§4) and its impact on the downstream NLI task by studying the following questions:

- **RQ1:** What is the efficacy of unsupervised approaches for evidence extraction? (§5.2)

[9] https://www.sbert.net

[10] Tabular probing using row deletion, row-value updation, row permutation, and row insertion.

[11] We explored other selection ratios too, take rows with rank till $5\times$, $2\times$, and $4\times$, but discovered that their performance is equivalent to (a), (b), and (c) respectively.

Category	Unsupervised Methods	α_1	α_2	α_3
Baseline	WMD (Gupta et al., 2020)	29.42	30.13	28.23
	DRR (Neeraja et al., 2021)	33.36	35.72	33.38
Static Embedding	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	71.49	73.28	63.41
Alignment	SimAlign (Match (mwmf))	58.98	61.53	66.33
Contextualised Embedding	Sentence-Transformer (paraphrase-mpnet-base-v2)	67.37	69.88	63.36
	SimCSE-Unsupervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$)	72.93	70.88	66.33
	SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$)	**75.79**	**75.74**	**68.81**
Human	Oracle (Gupta et al., 2021)	**88.62**	**89.23**	**88.56**

Table 2: F1-scores of the unsupervised evidence extraction methods.

- **RQ2:** Is supervision beneficial? Is it helpful to use hard negatives from unsupervised approaches for supervised training? (§5.2).
- **RQ3:** Does evidence extraction enhance the downstream tabular inference task? (§5.3)

5.1 Experimental Setup

First, we briefly summarize the models used in our experiments. We investigate both unsupervised (§4.1) and supervised (§4.2) evidence extraction methods. We use only the extracted evidence as the premise for the tabular inference task (§4.3). We compare both tasks against human performance.

As baselines, we use the **W**ord **M**over **D**istance (WMD) of Gupta et al. (2020) and the original DRR (Neeraja et al., 2021) with Top-4 extracted evidence rows. For DRR (Re-Rank + Top-S$^{\tau}$), which uses static embeddings, we set the sparsity parameter $S = 2$, and the dynamic row selection parameter $\tau = 1.0$. Our choice of S is based on the observation that in INFOTABS most (92%) instances have only one (54%) or two (38%) relevant rows. We set δ to 0.5 for all experiments.

For the Sentence Transformer, we used the *paraphrase-mpnet-base v2* model (Reimers and Gurevych, 2019) which is a pre-trained with the *mpnet-base* architecture using several existing paraphrase datasets. This choice is based on performance on the development set.

Both the supervised and unsupervised SimCSE models use the same parameters as DRR (Re-Rank + Top-K$_{\tau}$). We refer to the supervised and unsupervised variants as SimCSE-Supervised and SimCSE-Unsupervised respectively.

For the NLI task, we use the BPR representation over extracted evidence T^{R} with the RoBERTa$_{\text{LARGE}}$ two sentence classification model. We compare the following settings: (a) WMD Top-3 from Gupta et al. (2020), (b) No extraction i.e. using the full premise table with the "para" representation from Gupta et al. (2020), (c) DRR Top-4, (d) DRR (Re-Rank + Top-2$_{(\tau=1)}$) for training, development and test sets, (e) training a supervised classifier with a human oracle i.e. annotated evidence extraction as discussed in §3, and using the best extraction model, i.e. supervised evidence extraction with Hard Negative (3×) for the test sets, and (f) the human oracle across the training, development, and test sets.

5.2 Results of Evidence Extraction

Unsupervised evidence extraction. For RQ1, Table 2 shows the performance of unsupervised methods. We see that the contextual embedding method, SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$), performs the best. Among the static embedding cases, DRR (Re-Rank + Top-2$_{(\tau=1)}$) sees substantial performance improvement over the original DRR baseline. The alignment based approach using SimAlign underperforms, especially on the α_1 and α_2 test sets. However, its performance on the α_3 data, with out of domain and longer tables, is competitive to other methods.

Overall, the idea of using *Top-S_{τ}*, i.e., using the dynamic number of rows prediction and *Re-Rank* (exact-match based re-ranking) is beneficial. Previously used approaches such as DRR and WMD have low F1-score, because of poor precision. Using *Re-Rank* based on exact match improves the evidence extraction recall. Furthermore, introducing sparsity with *Top-S_{τ}*, i.e. considering only the Top-2 rows (S=2) and dynamic row selection ($\tau = 1$) substantially enhances evidence extraction precision. Furthermore, the zero weighting of title matches using the Hypo-Title-Swap heuristic, benefits contextualized embedding models such as SimCSE[12].

SimCSE-supervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$) outperforms DRR (Re-Rank + Top-2$_{(\tau=1)}$) by 4.3% (α_1), 2.5% (α_2) and 5.4% (α_3) absolute score. Since the table domains and the NLI reasoning involved for α_1 and α_2 are sim-

[12] For static embedding models, the effect of Hypo-Title-Swap was insignificant

Category	Evidence Extraction Train Set	Evidence Extraction Test Set	α_1	α_2	α_3
Baseline	WMD (Gupta et al., 2020)	WMD (Gupta et al., 2020)	70.38	62.55	61.33
	No Extraction (Gupta et al., 2020)	No Extraction (Gupta et al., 2020)	74.88	65.55	64.94
	DRR (Neeraja et al., 2021)	DRR (Neeraja et al., 2021)	75.78	67.22	64.88
Unsupervised	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	74.66	67.38	65.83
Supervised	Oracle	Supervised (3× Hard Negative)	77.34	71.15	68.92
Human	Oracle	Oracle (Gupta et al., 2021)	**78.83**	**71.61**	**71.55**
Human	Human NLI (Gupta et al., 2020)	Human NLI (Gupta et al., 2020)	**84.04**	**83.88**	**79.33**

Table 3: Tabular NLI performance with the extracted relevant rows as the premise.

ilar, so is their evidence extraction performance. However, the performance of α_3, which contains out-of-domain and longer tables (an average of thirteen rows, versus nine rows in α_1 and α_2) is relatively worse. The unsupervised approaches are still 12.69% (α_1), 13.49% (α_2), and 19.81% (α_3) behind the human performance, highlighting the challenges of the task.

Supervised evidence extraction. For RQ2, Table 4 shows the performance of the supervised relevant row extraction approaches that use binary classifiers trained with several sampling techniques for irrelevant rows. Overall, adding supervision is advantageous[13]. Furthermore, we observe that using the unsupervised DRR technique to extract challenging irrelevant rows, i.e., Hard Negative, is more effective than random sampling. Indeed, using random negative examples as the irrelevant rows performs the worst. Not sampling (6×) or using only one irrelevant row, namely Hard Negative (1×), also underperforms. We see that employing moderate sampling, i.e., Hard Negative (3×), performs best across all test sets.

The best supervised model with Hard Negative (3×) sampling improves evidence extraction performance by 8.7% (α_1), 10.8% (α_2), and 4.2% (α_3) absolute score over the best unsupervised model, namely SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-2$_{(\tau=1)}$). [14] The human oracle outperforms the best supervised model by 4.13% (α_1) and 2.65% (α_2) absolute scores—a smaller gap than the best unsupervised approach. We also observe that the supervision does not benefit the α_3 set much, where the performance gap to humans is still about 15.95% (only 3.80% improvement over unsupervised approach). We suspect this is because of the distributional changes in α_3 set noted earlier. This highlights directions for future improvement via domain adaptation.

[13] We investigate "How much supervision is adequate?" in Appendix A. [14] Although α_2 is adversarial owing to label flipping, rendering the NLI task more difficult, both α_1 and α_2 have instances with the same domain tables and hypotheses with similar reasoning types, making the relevant row extraction task equally challenging.

Sampling (Ratio)	α_1	α_2	α_3
Random Negative (1×)	69.42	71.94	54.12
Hard Negative (1×)	80.88	84.37	68.28
No Sampling (6×)	83.76	85.41	71.26
Hard Negative (3×)	**84.49**	**86.58**	**72.61**
Human Oracle	**88.62**	**89.23**	**88.56**

Table 4: F1-scores of supervised evidence extractors.

5.3 Results of Natural Language Inference

For RQ3, we investigate how using only extracted evidence as a premise impacts the performance of the tabular NLI task. Table 3 shows the results. Compared to the baseline DRR, our unsupervised DRR (Re-Rank + Top-2$_{(\tau=1)}$) performs similarly for α_2, worse by 1.12% on α_1, and outperforms by 0.95% on α_3.

Using evidence extraction with the best supervised model, Hard Negative (3×), trained on human-extracted (Oracle) rows results in 2.68% (α_1), 3.93% (α_2), and 4.04% (α_3) improvements against DRR. Furthermore, using human extracted (Oracle) rows for both training and testing sets outperforms all models-based extraction methods. The human oracle based evidence extraction leads to largest performance improvements of 3.05% (α_1), 4.39% (α_2), and 6.67% (α_3) over DRR. Overall, these findings indicate that extracting evidence is beneficial for reasoning in tabular inference task.

Despite using human extracted (Oracle) rows for both training and testing, the NLI model still falls far behind human reasoning (Human NLI) (Gupta et al., 2020). This gap exists because, in addition to extracting evidence, the INFOTABS hypotheses require inference with the evidence involving common-sense and knowledge, which the NLI component does not adequately perform.

6 Evidence Extraction: Human versus Model

We perform an error analysis of how well our proposed supervised extraction model (Hard Negative(3x)) performs compared to the human annotators. The model makes two types of errors: a Type I error occurs when an evidence row is marked as irrelevant, whereas Type II error occurs when an irrelevant row is marked as evidence. A Type I error will reduce the model's precision for the extraction model, whereas a Type II error will decrease the model's recall. Type I errors are especially concerning for the downstream NLI task. Since mislabeled evidence rows will be absent from the extracted premise, necessary evidence will be omitted, leading to inaccurate entailment labels. On the other hand, with Type II errors, when an irrelevant row is labeled as evidence, the model has to deal with from extra noise in the premise. However, all the required evidence remains.

Table 5 shows a comparison of the supervised extraction (Hard Negative (3x)) approach with the ground truth human labels on all the three test sets for both error types. On the α_3 set, Type-I and Type-II errors are substantially higher than α_1 and α_2. This highlights the fact that on the α_3 set, the model disagrees with with humans the most. Furthermore, the ratio of Type-II over Type-I errors is much higher for α_3. This indicates that the supervised extraction model marks many irrelevant rows as evidence (Type-II error) for α_3 set. The out-of-domain origin of α_3 tables, as well as their larger size, might be one explanation for this poor performance. Appendix §C provides several examples of both types of errors.

Test Set	Type-I	Type-II	Ratio (II/I)	Total
α_1	312	430	1.38	742
α_2	286	358	1.25	644
α_3	508	1053	2.07	1561

Table 5: Type-I and Type-II error of best supervised evidence extraction model.

7 Discussion

Why Sequential Prediction? Our choice of the sequential paradigm is motivated by the observation that it enforces a causal structure. Of course, a joint or a multi-task model may make better predictions. However, these models ignore the causal relationship between evidence selection and label prediction (Herzig et al., 2021; Paranjape et al., 2020). Ideally, each row is independent and, its relevance to the hypothesis can be determined on its own. In a joint or a multi-task model that exploits correlations across rows and the final label, *irrelevant rows* and the *NLI label*, can erroneously influence row selection.

Future Directions. Based on the observations and discussions, we identify the future directions as follows. (1) *Joint Causal Model.* To build a joint or a multi-task model that follows the causal reasoning structure, significant changes in model architecture are required. Such a model would first identify important rows and then use them for NLI predictions, but without risking spurious correlations. (2) *How much Supervision is Needed?* As evident from our experiments, relevant row supervision improves the evidence extraction, especially on α_1 and α_2 sets compared to unsupervised extraction. But do we need full supervision for all examples? Is there any lower limit to supervision? We partially answered this question in the affirmative by training the evidence extraction model with limited supervision (semi-supervised setting), but a deeper analysis is needed to understand the limits. See Appendix A for details. (3) *Improving Zero-shot Domain Performance.* As evident from §5.2, the evidence extraction performance of out-of-domain tables in α_3 needs further improvements, setting up a domain adaptation research question as future work. (4) Finally, inspired by Neeraja et al. (2021), we may be able to add explicit knowledge to improve evidence extraction.

8 Comparison with Related Work

Tabular Reasoning Many recent studies investigate various NLP tasks on semi-structured tabular data, including tabular NLI and fact verification (Chen et al., 2020b; Gupta et al., 2020; Zhang and Balog, 2019), various question answering and semantic parsing tasks (Zhang and Balog, 2020; Zhang et al., 2020b; Pasupat and Liang, 2015; Krishnamurthy et al., 2017; Abbas et al., 2016; Sun et al., 2016; Chen et al., 2020c; Lin et al., 2020; Zayats et al., 2021; Oguz et al., 2020; Chen et al., 2021, *inter alia*), and table-to-text generation (e.g., Parikh et al., 2020; Li et al., 2021; Nan et al., 2021; Yoran et al., 2021; Chen et al., 2020a).

Several strategies for representing Wikipedia relational tables are proposed, such as Table2vec (Deng et al., 2019), TAPAS (Herzig et al., 2020), TaBERT (Yin et al., 2020), TabStruc (Zhang

et al., 2020a), TABBIE (Iida et al., 2021), TabGCN (Pramanick and Bhattacharya, 2021) and RCI (Glass et al., 2021). Yu et al. (2018, 2021); Eisenschlos et al. (2020) and Neeraja et al. (2021) study pre-training for improving tabular inference.

Interpretability and Explainability Model interpretability can either be through explanations or by identifying the evidence for the predictions (Feng et al., 2018; Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; DeYoung et al., 2020; Paranjape et al., 2020). Additionally, NLI models (e.g. Ribeiro et al., 2016, 2018a,b; Zhao et al., 2018; Iyyer et al., 2018; Glockner et al., 2018; Naik et al., 2018; McCoy et al., 2019; Nie et al., 2019; Liu et al., 2019a) must be subjected to numerous test sets with adversarial settings. These settings can focus on various aspects of reasoning, such as perturbed premises for evidence selection (Gupta et al., 2021), zero-shot transferability (α_3), counterfactual premises (Jain et al., 2021), and contrasting hypotheses α_2. Recently, Kumar and Talukdar (2020) introduced Natural-language Inference over Label-specific Explanations (NILE), an NLI approach for generating labels and accompanying faithful explanations using auto-generated label-specific natural language explanations. Our work focuses on the extraction of label-independent evidence for correct inference, rather than on the generation of abstractive explanations for a given label.

Comparison with Shared Tasks The SemEval'21 Task 9 (Wang et al., 2021b) and FEVEROUS'21 shared task (Aly et al., 2021) are conceptually close to this work.

The SemEval task focuses on statement verification and evidence extraction using relational tables from scientific articles. In this work, we focus on item evidence extraction for non-scientific Wikipedia Infobox entity tables, proposed a two-stage sequential approach, and used the INFOTABS dataset which has complex reasoning and multiple adversarial tests for robust evaluation.

The FEVEROUS'21 shared task focuses on verifying information using unstructured and structured evidence from open-domain Wikipedia. Our approach concerns evidence extraction from a single table rather than open-domain document, table or paragraph retrieval. Furthermore, we are only concerned with entity tables rather than relational tables or unstructured text, while the FEVEROUS data has relational tables, unstructured text, and fewer entity tables.

9 Conclusion and Future Work

In this paper, we introduced the problem of *Trustworthy Tabular Inference*, where a reasoning model both extracts evidence from a table and predicts an inference label. We studied a two-stage approach, comprising an evidence extraction and an inference stage. We explored several unsupervised and supervised strategies for evidence extraction, several of which outperformed prior benchmarks. Finally, we showed that by using only extracted evidence as the premise, our approach outperforms previous baselines on the downstream tabular inference task.

Acknowledgements

The authors thank Bloomberg's AI Engineering team, especially Ketevan Tsereteli, Anju Kambadur, and Amanda Stent for helpful feedback and directions. We also appreciate the useful feedback provided by Ellen Riloff and the Utah NLP group. Additionally, we appreciate the helpful inputs provided by Atreya Ghosal, Riyaz A. Bhat, Manish Srivastava, and Maneesh Singh. Vivek Gupta acknowledges support from Bloomberg's Data Science Ph.D. Fellowship. This work was supported in part by National Science Foundation grants #1801446 (SaTC) and #1822877 (Cyberlearning). Finally, we would like to express our gratitude to the reviewing team for their insightful comments.

References

Faheem Abbas, M. K. Malik, M. Rashid, and Rizwan Zafar. 2016. Wikiqa — a question answering system on wikipedia using freebase, dbpedia and infobox. *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 185–193.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021. Open question answering over tables and text. In *International Conference on Learning Representations*.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020b. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Davide Chicco. 2021. *Siamese Neural Networks: An Overview*, pages 73–94. Springer US, New York, NY.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Li Deng, Shuo Zhang, and Krisztian Balog. 2019. Table2vec: Neural word and entity embeddings for table population and retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1029–1032, New York, NY, USA. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avi Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. 2021. Capturing row and column semantics in transformer based question answering over tables. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1212–1224, Online. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Vivek Gupta, Riyaz A. Bhat, Atreya Ghosal, Manish Srivastava, Maneesh Singh, and Vivek Srikumar. 2021. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *CoRR*, abs/2108.00578.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456, Online. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Nupur Jain, Vivek Gupta, Anshul Rai, and Gaurav Kumar. 2021. TabPert : An effective platform for tabular perturbation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 350–360, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, Copenhagen, Denmark. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Tongliang Li, Lei Fang, Jian-Guang Lou, and Zhoujun Li. 2021. TWT: Table with written text for controlled data-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1244–1254, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2020. Bridging textual and tabular data for cross-domain text-to-SQL semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4870–4888, Online. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

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

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

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

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2799–2809, Online. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of nli models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2020. Unified open-domain question answering with structured and unstructured knowledge. *arXiv preprint arXiv:2012.14610*.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Aniket Pramanick and Indrajit Bhattacharya. 2021. Joint learning of representations for web-tables, entities and types using graph convolutional network. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1197–1206, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings*

of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Huan Sun, Hao Ma, Xiaodong He, Scott Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the companion publication of the 25th international conference on World Wide Web*. ACM - Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021a. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021b. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Ori Yoran, Alon Talmor, and Jonathan Berant. 2021. Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills. *arXiv preprint arXiv:2107.07261. Version 1.*

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, bailin wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, richard socher, and Caiming Xiong. 2021. Gra{pp}a: Grammar-augmented pre-training for table semantic parsing. In *International Conference on Learning Representations*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. 2021. Representations for question answering from documents with tables and text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2895–2906, Online. Association for Computational Linguistics.

Hongzhi Zhang, Yingyao Wang, Sirui Wang, Xuezhi Cao, Fuzheng Zhang, and Zhongyuan Wang. 2020a. Table fact verification with structure-aware transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1629, Online. Association for Computational Linguistics.

Shuo Zhang and Krisztian Balog. 2019. Auto-completion for data cells in relational tables. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 761–770, New York, NY, USA. ACM.

Shuo Zhang and Krisztian Balog. 2020. Web table extraction, retrieval, and augmentation: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(2):13:1–13:35.

Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020b. Summarizing and exploring tabular data in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1537–1540, New York, NY, USA. Association for Computing Machinery.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

A How Much Supervision is Enough for Evidence Extraction?

To investigate this, we use Hard Negative (3x) with RoBERTa$_{LARGE}$ model as our evidence extraction classifier, which is similar to the full supervision method. To simulate semi-supervision settings, we randomly sample 10%, 20%, 30%, 40%, and 50% example instances of the train set in an incremental fashion for model training, where we repeat the random samplings three times. Figure 3, 4, and 5 compare the average F1-score over three runs on the three test sets α_1, α_2 and α_3 respectively.

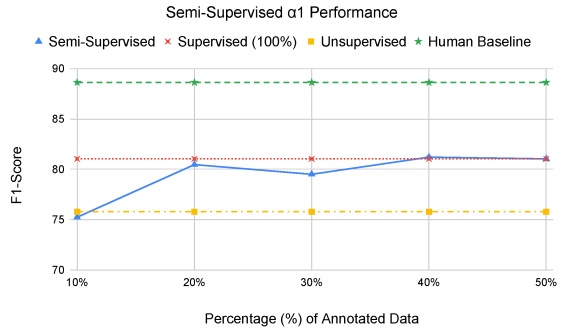

Figure 3: Extraction performance with limited supervision for α_1. All results are average of three random splits runs.

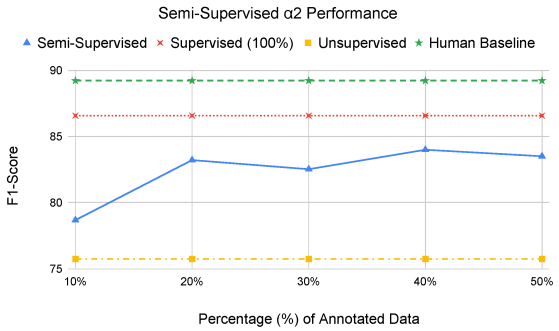

Figure 4: Extraction performance with limited supervision for α_2. All results are average of three random splits runs.

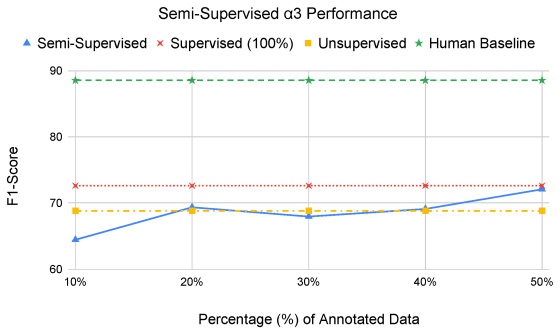

Figure 5: Extraction performance with limited supervision for α_3. All results are average of three random splits runs.

We discovered that adding *some* supervision had advantages over not having any supervision. However, we also find that 20% supervision is adequate for reasonably good evidence extraction with only $< 5\%$ F1-score gap with full supervision. One key issue we observe is the lack of a visible trend due to significant variation produced by random data sub-sampling. It would be worthwhile to explore if this volatility could be reduced by strategic sampling using an unsupervised extraction model, an active learning framework, and strategic diversity maximizing sampling, which is left as future work.

B Human Annotation Quality Control

Since many hypothesis sentences (especially those with neutral labels) require out-of-table information for inference, we introduced the option to choose out-of-table (OOT) pseudo rows, which are highlighted only when the hypothesis requires information that is not common (i.e. common sense) and missing from the table. To reduce any possible bias due to unintended associations between the NLI label and the row selections (e.g., using OOT for neutral examples), we avoid showing inference

labels to the annotators[15].

To assess an annotator, we compare their annotations with the majority consensus of other annotators' (four) annotations. We perform this comparison at two levels: (a) **local-consensus-score** on the most recent batch, and (b) **cumulative-consensus-score** on all batches annotated thus far.

We use these consensus scores to temporarily (local-consensus-score) or permanently (cumulative score) block the poor annotators from the task. We also review the annotations manually and provide feedback with more detailed instructions and personalized examples for annotators who were making mistakes due to ambiguity in the task. We give incentives to annotators who received high consensus scores. As in previous work, we removed certain annotators' annotations that have a poor consensus score (cumulative score) and published a second validation HIT to double-check each data point if necessary.

C Human vs Models Qualitative Examples

We manually inspect the Type I and Type II error instances for the supervised model and human annotation for the development set. Below, we show some of these examples where models conflict with ground-truth human annotation. We also provide a possible reason behind the model mistakes.

Type I. Below, we show Type I error examples.

Example I
Row: Colorado Springs, Colorado is a poor training location for endurance athletes.

Hypothesis: The elevation of Colorado Springs, Colorado is 6,035 ft (1,839 m).

Model Prediction: Not Relevant
Human Ground Truth: Relevant Evidence.

Possible Reason: Model wasn't able to connect the concept of elevation with the perfect high elevation training ground requirement of endurance athletes. Requires common sense and knowledge.

[15] Because of the random sequence and unbalanced nature, each of the three hypothesis sentences can have any NLI label, i.e., in total $3^3 = 27$ possibilities.

Example II
Row: The number of number of employees of International Fund for Animal Welfare - ifaw is 300+ (worldwide).

Hypothesis: International Fund for Animal Welfare - ifaw is a national organization focused on only North America.

Model Prediction: Not Relevant
Human Ground Truth: Relevant Evidence.

Possible Reason: Model wasn't able to connect the clue ('worldwide') in the table row with the phrase 'focused on only north America'.

Example III
Row: The equipment of Combined driving are horse, carriage, horse harness equipment.

Hypothesis: Combined driving is a horse racing event style.

Model Prediction: Not Relevant
Human Ground Truth: Relevant Evidence.

Possible Reason: Model wasn't able to connect the horse related equipment i.e. 'horse carriage, horse harness' with the event time i.e. 'horse racing'.

Type II. Below, we show Type II error examples.

Example I **Row**: Dazed and Confused was directed by Richard Linklater.

Hypothesis: Dazed and Confused was directed in 1993.

Model Prediction: Relevant Evidence
Human Ground Truth: Not Relevant.

Possible Reason: Model focuses on lexical match token 'directed' instead using entity type where premise refer for 'Person' who directed rather than 'Date' of direction.

Example II **Row**: The spouse(s) of Celine Dion (CC OQ ChLD) is René Angélil, (m. 1994; died 2016).

Hypothesis: Thérèse Tanguay Dion had a child that became a widow.

Model Prediction: Relevant Evidence
Human Ground Truth: Not Relevant.

Possible Reason: Model was unable to connect widow concept in hypothesis with it relation to Spouse and the marriage date René Angélil, (m. 1994; died 2016).

Example III **Row**: The trainer of Caveat is Woody Stephens.

Hypothesis: Caveat won more in winnings than it took to raise and train him.

Model Prediction: Relevant Evidence
Human Ground Truth: Not Relevant.

Possible Reason: Model connects the 'raise and train' term with the trainer name which is unrelated and has no connection to overall, winning races money vs spending for animal.

Discussion Based on the observation from the above examples as also stated in $5.3, the model fails on many examples due to its lack of knowledge and common-sense reasoning ability. One possible solution to mitigate this is by the addition of implicit and explicit knowledge on-the-fly for evidence extraction, as done for inference task by Neeraja et al. (2021).

D Implicit Relevance Indication

We manually examine the human-annotated evidence in the development set. We discovered the existence of several relevant phrases/tokens which implicitly indicate the presence of evidence rows. E.g. The existence of tokens such as *married*, *husband*, *lesbian*, and *wife* in hypothesis (H) is very suggestive of the row *Spouse* being the relevant evidence. Learning such implicit relevance-based phrases and tokens connection is easy for humans and large pre-trained supervision models. It is a challenging task for similarity-based unsupervised extraction methods. Below, we show implicit relevance, indicating token and the corresponding relevant evidence rows.

Relevance Indicating Phrase (H) → Relevant Evidence Rows Key(T)

'broked', 'started from', 'doesn't anymore', 'still perform', 'over a decade', 'began performing', 'started wrapping', 'first started' → year active

age related term, 'were of <age>', 'after <age>', 'fall', 'spring','birthday' → born

'several years', 'one month', century art → years

'co-wrote', 'written', 'writer', 'original written' → written by (novel and book)

'married', 'husband', 'lesbian', 'wives' → Spouse

'no-reward', 'monetary value', 'prize' → rewards

'earlier', 'debut', '21st century', 'early 90s', 'recording','product of years' → recorded

'lost', 'won', 'races','competition' → records (horse races, car races etc) 'sea level' → 'lowest elevation', 'highest elevation', 'elevation'

multi-lingual, multi-faith → 'regional languages', 'official languages', 'religion', ','race or faith'

'acting', 'rapping', 'politics' → occupation

'over an', 'shortest', 'longest', 'run-time' → length 'is form <country>', 'originate', 'are an <nationality>', 'formed on <location>', 'moved to <Country>', 'descended from' → origin, descendant, parenthood etc

'city' with 'x' peoples → 'metropolitan municipality' or 'metro'

'was painted with', 'mosaic', 'oil', 'water' → medium

'hung in' , 'museum', 'is stored in/at', 'wall', 'mural' → 'location'

'was discontinued', 'awards' → 'last awarded'

'playing bass' → 'instruments'

'served', 'term', 'current charge' , 'in-charge' → 'in office'

'is controlled by', 'under control' → 'government'

'classical', 'pop', 'rock', 'hip-hop', 'sufi' → genre

'won more', 'in winning (race)', 'earned more than' → earnings

'Register of', 'Cultural Properties' → designated

'urban area', 'less dense' -> urban density, density

'founded by', 'has been around', 'years' → founded , introduce

'was started', 'century', 'was formed', '100 years' → founded, formation

'daughters', 'sons' → children spouse(s), partner(s)

'lost money', 'net profit', 'budget', 'unprofitable', 'not popular'(common sense)

'owned' or 'company' → manufacturer

'bigger than an average' → dimension