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

Anonymous ACL submission

Abstract

When pre-trained contextualized embeddings-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 *ignore* or focus on *wrong* evidence to predict labels. 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. To begin, 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 the automatically extracted evidence as the only 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" represent the mapping of the hypothesis sentences with the evidence rows. The colored text (and subscript number) 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., 2019) 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 models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b). These models, when adapted for tabular NLI by flattening tables into synthetic sentences using heuristics, 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* the relevant rows to (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) use existing pre-trained knowledge (Jain et al., 2021; Gupta et al., 2021) for inference. In essence, the models use spurious correlations between irrelevant rows, the hypothesis and the inference label for predicting labels.

In this paper, we argue that existing NLI systems optimized solely for label prediction cannot be fully trusted. It is not sufficient for a model be merely *"Right"* but also *"Right for the Right Reasons"*. Thus, extraction of relevant rows as the *"Right Reasons"* is equally important for trustworthy reasoning[1]. We address this issue, by introducing

[1]We suggest that a reasoning system can be deemed trustworthy only if it exposes how its decisions are made, thus verifying whether it is right for the right reasons.

the task of *Trustworthy Tabular Inference*, where the goal is to focus on both extracting relevant rows as evidence and predicting inference labels.

To illustrate this task, let us look at an example from the INFOTABS dataset in Figure 1, which shows a premise table and three hypotheses. This example also depicts the evidence rows and the corresponding tokens in hypothesis that indicates the relevance connection link. For trustworthy tabular reasoning, the model, in addition to predicting label ENTAIL for *H1*, CONTRADICT for *H2* and NEUTRAL for *H3*, also identifies the evidence rows. i.e., rows *Producer* and *Length* for hypothesis *H1*, *Recorded* for hypothesis *H2*, *Released* and *Recorded* for hypothesis *H3*.

We propose a two-stage sequential prediction approach, which comprises of an evidence extraction stage and the inference stage. In the evidence extraction stage, the model focuses on extracting the necessary evidence information needed for reasoning. During inference stage, the NLI model then uses only the extracted evidence as the premise for label prediction task.

We explore several unsupervised evidence extraction approaches on 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 annotated the INFOTABS training set (17K table-hypothesis pairs with 1740 unique tables) with relevant rows following Gupta et al. (2021), and then train a RoBERTa$_{Large}$ classifier. The supervised model further enhances the evidence extraction performance by 8.7%, 10.8%, and 4.2% absolute score on the three test sets over unsupervised approaches. Finally, for the full inference task, we demonstrate that our two-stage approach with best extraction, outperform the earlier baseline by 1.6%, 3.8%, and 4.2% absolute score on the three test sets.

In summary, our contributions are as follows:

- We introduce the problem of trustworthy tabular reasoning and propose a two-stage prediction approach that includes an evidence extraction stage and an inference stage.
- We investigate a variety of unsupervised evidence extraction techniques. Our unsupervised approach for evidence extraction outperform the previous methods.
- We enrich the INFOTABS train set with evidence rows and develop a supervised extraction approach with human-like performance.
- We demonstrate that our two-stage technique with best extraction outperforms all the prior benchmarks on the downstream NLI task.

The updated dataset, along with associated code, is available at `anonymous_for_submission`.

2 Task Formulation

We begin by introducing the task formulation and datasets we are working on.

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, this task maps them to ENTAIL (E), CONTRADICT (C) or NEUTRAL (N) as

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

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

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

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*.

Dataset Details. There are several datasets for tabular NLI: TabFact, INFOTABS, and the SemEval'21 Task 9 (Ru Wang et al., 2021) and the FEVEROUS'21 shared task (Aly et al., 2021) datasets. We use the INFOTABS data. It contains finer-grained annotation (e.g., TabFact lacks NEUTRAL hypotheses) and complex reasoning[2] than the others.

[2] As per Gupta et al. (2020), examples in INFOTABS require complex reasoning involving multiple rows (33%). The dataset covers all reasoning types present in Glue (Wang et al., 2018) and SuperGlue (Wang et al., 2019).

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

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

The dataset consists of $23,738$ premise-hypothesis pairs collected by 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 associated NLI label. All tables contain a *title* followed two columns (cf. Figure 1, left columns are *keys* and right are *values*).

In addition to the train and dev sets, the data includes multiple adversarial test sets: 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; 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 Evidence Extraction by Human

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 3 examples. To reduce bias induced by a link between the NLI label and row selections, we do not provide labels to the annotators. The quality control details are provided in the Appendix A.

In total, we received 81,282 annotations from 90 distinct annotators. Overall, twenty five annotators completed more than 1000 tasks, corresponding to 87.75 % examples, indicating a tail distribution with the annotations. In the end, 16,248 training set table-hypothesis pairs were successfully labeled with the evidence rows[3]. 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[4].

[3] 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.

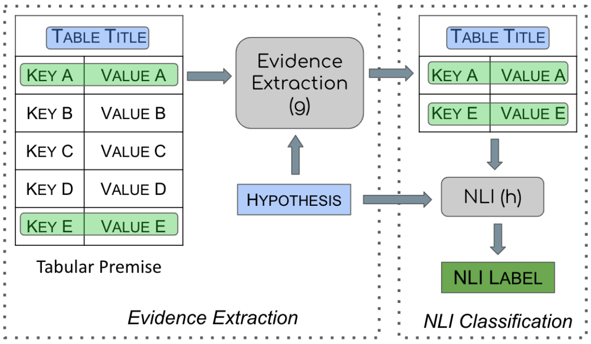

Figure 2: High level flowchart showing our approach for evidence extraction and trustworthy tabular inference.

Choice of Semi-structured Data. Despite connection to title entity, the table's rows are semantically distinct. Each row can be considered as 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 true for unstructured text, where granularity at the token, phrase, and paragraph levels is missing (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)[5]. To operationalize this intuition, we chose a two-stage sequential approach which consists of an evidence extraction followed by the NLI classification, 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

[4] We also manually examined hypothesis phrases that signal relevant rows. See Appendix E for details.

[5] See more details discussion in section 6

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

Here, the function 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 $g(.)$, $h(.)$ 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 INFOTABS.

We explore unsupervised (Section 4.1) and supervised (Section 4.2) evidence extraction methods to model the function $g(.)$, i.e., the evidence row extraction.

4.1 Unsupervised Evidence Extraction

All the unsupervised approaches extract Top-K rows based on relevance scores, where K is a hyper-parameter. To score rows, we use the cosine similarity between the row and the hypothesis sentence representations. We study three categories of evidence extraction methods, as described below.

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. To improve DRR technique, we proposed three modifications as follows.

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

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 a fixed K rows. If the similarity (after Re-Rank) between the row and the hypothesis sentence representations $>$ threshold (τ) we select the row otherwise not.

We adapt this strategy because: (a) The number of rows in premise table can vary across examples, (b) and hypothesis can require different number of evidence rows for reasoning.

4.1.2 Using Embedding Alignment

This approach constitutes of two parts (a) getting alignment between rows and the hypothesis words (b) , and then computing cosine similarity between the alignment words. In specific, we use SimAlign (Jalili Sabet et al., 2020) method for getting word-level *alignment*. SimAlign use static and contextualized embeddings without parallel training data for getting words alignment. We choice the **Match (mwmf)** method for alignment matching. Match method uses **m**aximum-**w**eight **m**aximal matching **(mwmf)** in the bipartite weighted network formed by the word level similarity matrix (e.g., (Kuhn, 2010)), and finds a global optima. We prefer Match (mwmf) over the other greedy methods Itermax and Argmax because they finds only local optima. After alignment, we normalize the sum of cosine similarities of RoBERTa$_{Large}$ token embeddings[6] to derive the *relevance score*. Furthermore, because all rows use the same title, we assign title matching terms zero weight. We refer this method as SimAlign (Match (mwmf)) in this paper.

4.1.3 Using Contextualised Embeddings

Methods in Section 4.1.2 only provide alignment between words, but here, we compute similarity scores directly between the contextualised sentence embeddings obtained by transformer models. We explore two options here.

Sentence Transformer: We use Sentence-BERT (Reimers et al., 2019) and its variants (Reimers and Gurevych, 2020; Thakur et al., 2021; Wang et al., 2021). These model uses the Siamese neural network (Koch et al., 2015; Chicco, 2021) based loss objective. We explore several pre-trained sentence

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

transformers models[7] for sentence representation. These model differ in (a) the data used for pre-training (b) , the main model type and it size (c) , and, the maximum sequence length.

SimCSE: SimCSE (Gao et al., 2021) use a simple contrastive learning framework to train sentences embeddings in both unsupervised and supervised settings. The former takes an input sentence and predicts itself using standard dropout as the noise, and the latter takes example pairs from the MNLI dataset with *entailment* serving as *positives* and *contradiction* serving as *hard negatives* for contrastive learning.

We pass the rows sentence directly to SimCSE to get embeddings. Since all rows uses the same title, to avoid spurious matching between the hypothesis tokens and premise rows title tokens, we swap the hypothesis title tokens with another title (prefer single token title) from another table of same category (randomly selected). We then use the cosine similarity between SimCSE sentences embeddings to compute final relevance score. We again uses the sparsity and Dynamic selection as earlier. In the study, we refer 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 (Section 3) to construct supervised extraction training dataset. It contains hypothesis and each table-row with an binary label, signifying whether the row is relevant or irrelevant, obtain by human annotations. We use the sentences from Better Paragraph Representation (BPR) (Neeraja et al., 2021) to represent each row.

Label Balancing. The number of irrelevant rows would be substantially more than that of relevant rows for a table-hypothesis pair. It was empirically confirmed through our annotation analysis and independently by Gupta et al. (2021) through perturbation probing. Therefore, if we use all irrelevant rows from tables as negative examples, the resulting training set would be highly imbalanced, with about $6\times$ more irrelevant than relevant rows.

We investigate several *label balancing* strategies by *sub-sampling* the number of irrelevant rows for training. We explore the following schemes: (a) Take all irrelevant rows from the table without sub-sampling (on average $6\times$ more irrelevant rows) referred as **Without Sample**($6\times$), (b) pick unrelated rows at random in the same proportion as relevant rows referred as **Random Negative**($1\times$), (c) use the unsupervised DRR (Re-Rank + Top-S$_{\tau}$) method to pick the most irrelevant row in equal proportion as the relevant rows, referred as **Hard Negative**($1\times$), and (d) same to (c), except pick top three irrelevant rows, referred as **Hard Negative**($3\times$)[8].

Classifier Training. We use RoBERTa$_{Large}$ two sentence classifier for modeling the relevant-vs-irrelevant row classification. We prefer RoBERTa$_{Large}$, because of (a) superior performance in comparison to other models, and (b) the fact that RoBERTa$_{Large}$ is also used by Gupta et al. (2020); Neeraja et al. (2021) for the NLI task.

4.3 Natural Language Inference

For the downstream NLI task, the function $h(.)$ is a two-sentence classifier with input T^R (the output of the function $g(.)$) and hypothesis H. We use BPR for representing T^R as we did for T. Since |T^R| $\ll$ |T|, the extraction benefits larger tables (especially in α_3 set) which exceed the classifier token limit.

5 Experimental Evaluation

Our experiments assess the efficacy of evidence extraction (Section 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? (Section 5.2)

RQ2: Is supervision beneficial? Is it helpful to use hard negatives from unsupervised approaches for supervised training? (Section 5.2).

RQ3: Does evidence extraction enhance the downstream tabular inference task? (Section 5.3)

5.1 Experimental Setup

Next, we discuss the models used for experiments.

We investigate both unsupervised (Section 4.1) and supervised (Section 4.2) evidence extraction methods. Furthermore, we use the extracted evidence as the only premise for tabular inference task

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

[8] 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.

(Section 4.3). We compare both tasks with human performance.

As baselines, we use the **W**ord **M**over **D**istance (WMD) of Gupta et al. (2020) and the un-changed 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 maximum sparsity parameter S = 2, and the dynamic row selection parameter τ = 1.0. For simplicity and a fair comparison, we maintain δ at a constant 0.5 for all approaches. We choose S = 2, because in INFOTABS most (92%) instances have only one (54%) or two (38%) relevant rows.

As for models using *contextualized embeddings*, for the the Sentence Transformer, we used the *"paraphrase-mpnet-base v2"* model (Reimers et al., 2019) which is a pre-trained with the *mpnet-base* architecture using several exiting paraphrase/non-paraphrase datasets. Our choice of the *"paraphrase-mpnet-base v2"* model was guided by performance on the dev set. SimCSE (Gao et al., 2021) (both Supervised / Unsupervised) models uses the same parameters as **DRR (Re-Rank + Top-K$_{\tau}$)**. In addition, we use Hypo-Title-Swap to mitigate spurious matches from matching the title. We refer to the supervised and unsupervised variants as SimCSE-Supervised and SimCSE-Unsupervised.

For the NLI task we use the BPR representation on extracted evidence T^{R} with the RoBERTa$_{Large}$ two sentence classification model. We compare (a) Gupta et al. (2020) WMD Top-3, (b) No Extraction i.e. full premise table as "para" representation 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 Section 3, and using the best extraction model, i.e. supervised evidence extraction with Hard Negative (3×) for the test sets, (f) and, the human oracle across the training, development and the test sets.

5.2 Results of Evidence Extraction

Unsupervised: With regard to 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, SimAlign, performs worse, especially on the α_1 and α_2 test sets. However, surprisingly, 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. Prior models such as DRR and WMD have very low F1-score, because of poor precision. Using *Re-Rank* based on exact match improves the evidence extraction recall. Furthermore, introducing sparsity *Top-S_{τ}*, i.e. considering only the Top-2 rows (S=2) and dynamic row selection ($\tau = 1$) substantially enhance evidence extraction precision. Furthermore, the zero weighting of title matches, a.k.a Hypo-Title-Swap, benefits contextualized embedding model such as SimCSE[9].

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 similar, 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 comparatively 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: With regard to RQ2, Table 4 shows the performance of the supervised relevant row extraction using binary classification with several sampling techniques for irrelevant rows. Overall, adding supervision is advantageous[10]. Furthermore, we observe that using the unsupervised DRR technique to extract challenging irrelevant rows, a.k.a Hard Negative, is more effective[11] than random sampling. Indeed, using random negative examples as the irrelevant row performs the worst. Not sampling (6×) or using only one irrelevant row, namely Hard Negative (1×), also performs poorly. We see that employing moderate sampling, i.e., Hard Negative (3×) performs best.

The best supervised model with Hard Negative (3×) sampling enhanced evidence extraction per-

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

[10]To investigate "How much supervision is adequate?" we provide details in Appendix B

[11]Similar recall for DRR and SimCSE for Top-4 rows.

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
	Sentence-Transformer (paraphrase-mpnet-base-v2)	67.37	69.88	63.36
Contextualised Embedding	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-Score for several unsupervised evidence extraction method.

Category	Evidence Extraction Train Set	Evidence Extraction Test Set	α_1	α_2	α_3
	WMD (Gupta et al., 2020)	WMD (Gupta et al., 2020)	70.38	62.55	61.33
Baseline	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 (3x 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.

formance further by 8.7% (α_1), 10.8% (α_2), and 4.2%α_3 absolute score in comparison to best unsupervised model's evidence extraction, i.e., SimCSE-Supervised (Hypo-Title-Swap + Re-Rank + Top-$2_{(\tau=1)}$). The human oracle outperforms the best supervised model by 4.13%(α_1) and 2.65% (α_2) absolute scores, which is a smaller gap compared to the best unsupervised approach. Furthermore, we observe that the supervision does not benefit the α_3 set much, where the performance gap with human reduction is still around 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 future improvement directions by domain adaptation for supervised methods. Appendices C and D show more detailed error analysis for the interested reader.

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-Score for several supervised evidence extraction method. Here, (*) represent the human selected optimal rows.

5.3 Results of Natural Language Inference

For RQ3, we investigate how using only extracted evidence as premise impacts the performance of the tabular NLI task. Table 3 shows the results. In comparison 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 Discussion

Why Sequential Stages? 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 can make the predictions even better. However, this technique risks failing to fulfill the causal relationship between evidence selection and label prediction (Herzig et al., 2021; Paranjape et al., 2020). Ideally, each row is independent and determines the relevance to the hypothesis on its own. However, in a joint or a multi-task model that promotes spurious correlation, *irrelevant rows* and *NLI label*, can erroneously influence row selection (Gupta et al., 2021).

Future Directions. Based on the observations and discussions, we identify the future directions as follows. (a) *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; the model first latently identifies important rows and then uses them for NLI predictions. (b) *How much Supervision is Needed?* As evident from our experiments, relevant rows 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? Probably yes, we partially answered this question by training the evidence extraction model with limited supervision (semi-supervised setting); see Appendix B for details. (c) *Improving Zero-shot Domain Performance.* As evident from section 5.2, the evidence extraction performance of out-of-domain tables in α_3 can be further improved by transfer learning for domain adaptations, and (d) Lastly, inspired from (Neeraja et al., 2021), one can add implicit or explicit knowledge to improve evidence extraction, as evident from the error analysis in Appendix D.

7 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., 2019; Gupta et al., 2020), various question answering and semantic parsing tasks (Pasupat and Liang, 2015; Krishnamurthy et al., 2017; Abbas et al., 2016; Sun et al., 2016; Chen et al., 2020b; 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; Radev et al., 2020; Yoran et al., 2021; Chen et al., 2020a).

Several strategies for representing Wikipedia relational tables were recently proposed, such as TAPAS (Herzig et al., 2020), TaBERT (Yin et al., 2020), TabStruc (Zhang et al., 2020), 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 referring to 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.

Comparison with Shared Tasks The most closest work to our approach is the SemEval'21 Task 9 (Ru Wang et al., 2021) and FEVEROUS'21 shared task (Aly et al., 2021). SemEval focuses on statement verification and evidence finding using relational tables from scientific articles. Compared to SemEval, we focus on (a) evidence extraction for non-scientific Wikipedia Infobox entity tables, (b) proposed two stages sequential approach which follows casual reasoning aspect, (c) use 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 is more concerned on evidence extraction from a single table rather than open-domain document/table/paragraph retrieval. Furthermore, we are only concerned with entity tables rather than relational tables or unstructured text[12].

8 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 inference stage. We explored several unsupervised and supervised strategies for evidence extraction, several of which outperform prior benchmarks. Finally, we showed that using only extracted evidence as to the premise, our inference stage can outperform previous baselines at tabular inference.

[12]FEVEROUS has relational tables, unstructured text, and fewer entity tables

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 Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

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 Wang, and William W Cohen. 2021. Open question answering over tables and text. *Proceedings of ICLR 2021*.

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.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. 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 Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *Findings of EMNLP 2020*.

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.

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 Mueller. 2020. Understanding tables with intermediate pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 281–296.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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.

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.

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.

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.

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.

Harold W Kuhn. 2010. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008*, pages 29–47. Springer.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2020. Bridging textual and tabular data for cross-domain text-to-sql semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4870–4888.

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.

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.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of nli models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 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.

Ankur P 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 EMNLP*.

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.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *NAACL HLT 2018*, page 180.

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.

Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Nazneen Fatema Rajani, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2020. Dart: Open-domain structured data record to text generation. *arXiv preprint arXiv:2007.02871*.

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.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 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*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

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.

Nancy Xin Ru Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. Semeval2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (semtab-facts). *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.

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.

Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 771–782.

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. 2021. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*.

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.

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.

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. *arXiv preprint arXiv:2106.09449*.

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*.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2021. Grappa: Grammar-augmented pre-training for table semantic parsing. *International Conference of Learning Representation*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 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.

Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. 2021. Representations for question answering from documents with tables and text. *arXiv preprint arXiv:2101.10573*.

Hongzhi Zhang, Yingyao Wang, Sirui Wang, Xuezhi Cao, Fuzheng Zhang, and Zhongyuan Wang. 2020. 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.

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

A Quality Control for Crowdsourcing Evidence Extraction

Since many hypothesis sentences (especially those with neutral labels) require out-of-table information for inference, we've introduced the option to choose out-of-table (OOT) pseudo rows, which are highlighted only when the hypothesis requires information that isn't common (a.k.a common sense) and missing from the table. To reduce any possible bias due to unintended link between the NLI label and the row selections, e.g., using OOT for neutral examples, we avoid showing labels to the annotators[13].

To assess an annotator, we compare his/her annotation with the majority consensus of other annotators' (4) 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 spurious annotators from the task. We also review the annotations manually and provide feedback in term follow-up recommendations with more detailed instructions and personalized examples for genuine annotators who were making unkindly mistakes due to task uncertainty.

We give incentives to annotators who received high consensus scores, suggesting that they performed brilliantly on the assignment. As in previ-

[13] 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.

ous work, we remove certain annotators' annotations that have a very poor consensus score (cumulative score) and publish a second validation HIT to double-check each data point if necessary.

B 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 compares the average F1-score over three runs on the three test sets α_1, α_2 and α_3 respectively.

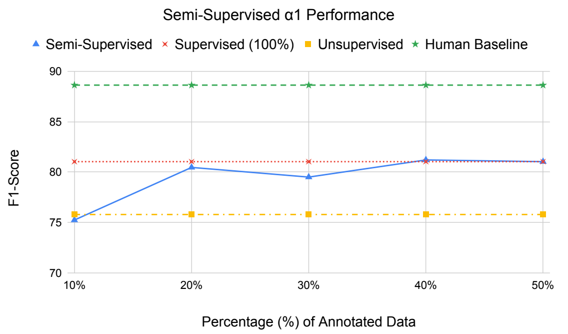

Figure 3: Extraction performance with limited supervision for α_1. All results are average of three random splits runs.

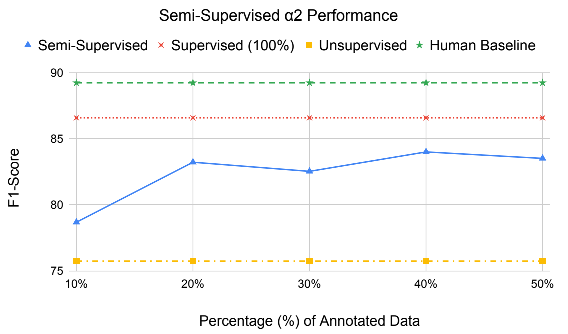

Figure 4: Extraction performance with limited supervision for α_2. All results are average of three random splits runs.

We discovered that adding supervision had advantages over not having any supervision. In addition, we find 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.

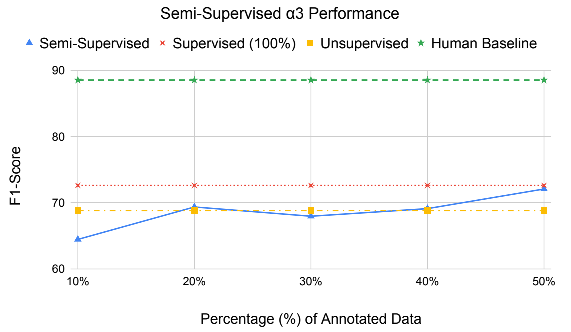

Figure 5: Extraction performance with limited supervision for α_3. All results are average of three random splits runs.

C Error Analysis: Human v.s. Supervised Models on Evidence Extraction

We perform an error analysis of how well does our proposed supervised extraction model (Hard Negative(3x)) performs as opposed to the human. The model makes two types of errors, referred to as Type I and Type II. Type I error occurs when an evidence row (1) is marked as non-relevant (0), whereas, Type II error occurs when an irrelevant row is marked as evidence. For the extraction model, a Type I error will reduce the model's precision, whereas a Type II error, will decrease the model's recall. The Type I mistake is especially concerning for the downstream NLI task; the mislabeled evidence rows (0 instead of 1) will be absent from the extracted premise, therefore necessary evidence will be omitted, resulting in inaccurate label prediction. On the other hand, in the Type II mistake, when an irrelevant row is labeled as evidence (1 instead of 0), the model just suffers from extra noise with the premise, but all required evidence remains.

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.

Table 5 shows a comparison of the supervised extraction (Hard Negative (3x)) approach with the provided ground truth human label for all the three test sets on both error types. On α_3 set the both Type-I and Type-II error is substantially higher than α_1 and α_2. This highlights that for the α_3 set the model has the worst disagreement with humans. Furthermore, the ratio of Type-II over Type-I error is substantially higher for α_3 than for α_1 and α_2. 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.

D Human vs Models Qualitative Examples

We manually inspect the Type I and Type II error examples 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 the possible reason behind the model mistakes.

Type I Error. 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. Require common sense and knowledge.

Example II
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'.

Example III
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'.

Type II Error. 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 focus 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 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 connect '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 Section 6.2 (d.), 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).

E Implicitly Relevance Indication Phrases

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 although easy for humans as well for large pre-trained supervision models, it is an incredibly difficult task for similarity-based unsupervised extraction methods. Below, we show implicit relevance indicating token and the corresponding relevant evidence rows.

Implicit Relevance Indicating Phrase (H) → Relevant Evidence Rows Keys (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 → year (painting category)

‘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)

’tall’, ’short’ → ’lowest’, ’highest’, ’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

‘owned’ or ‘company’ → manufacturer

‘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
‘founded by’, ‘has been around’, ‘years’ → founded , introduced

‘was started’, ‘century’, ‘was formed’, ’100 years’ → founded, formation

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

‘bigger than an average’ → dimension

‘Register of’, ‘Cultural Properties’ → designated

‘urban area’, ‘less dense’ -> urban density, density

‘American’, ‘British’, ‘European’, ‘from USA’ → country

‘daughters’, ‘sons’ → children spouse(s), partner(s)

‘is a bovine’(dog) → ’breed’

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