SemEval-2021 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS)

Nancy X. R. Wang* **Diwakar Mahajan*** **Marina Danilevsky** **Sara Rosenthal§**
IBM Research
nancywang1991@gmail.com, {dmahaja, mdanile, sjrosenthal}@us.ibm.com

Abstract

Understanding tables is an important and relevant task that involves understanding table structure as well as being able to compare and contrast information within cells. In this paper, we address this challenge by presenting a new dataset and tasks that addresses this goal in a shared task in SemEval 2020 Task 9: Fact Verification and Evidence Finding for Tabular Data in Scientific Documents (SEM-TAB-FACTS). Our dataset contains 981 manually-generated tables and an auto-generated dataset of 1980 tables providing over 180K statement and over 16M evidence annotations. SEM-TAB-FACTS featured two sub-tasks. In sub-task A, the goal was to determine if a statement is supported, refuted or unknown in relation to a table. In sub-task B, the focus was on identifying the specific cells of a table that provide evidence for the statement. 69 teams signed up to participate in the task with 19 successful submissions to subtask A and 12 successful submissions to subtask B. We present our results and main findings from the competition.

1 Introduction

Tables are ubiquitous in documents and presentations for conveying important information in a concise manner. This is true in many domains, stretching from scientific to government documents. In fact, surrounding text in these articles are often statements summarizing or highlighting some information derived from the primary source of data in tables. A relevant example is shown in Figure 1 from a Business Insider article analyzing the impact of Covid-19 (Aylin Woodward and Gal, 2020). Describing all the information provided in this table in a readable manner would be lengthy and considerably more difficult to understand. Despite their importance, popular question answering (e.g. SQuAD and Natural Question (Rajpurkar et al., 2016; Kwiatkowski et al., 2019)) and truth verification tasks (e.g. SemEval-2019 Fact Checking Task (Mihaylova et al., 2019)) have not focused on tables, being composed solely of written text. This is likely due to their complexity to parse and understand, despite their rich amount of information.

The total number of cases and deaths have far surpassed those of the SARS outbreak.

2019 novel coronavirus compared to other major viruses

VIRUS	YEAR IDENTIFIED	CASES	DEATHS	FATALITY RATE	NUMBER OF COUNTRIES
Ebola	1976	33,577	13,562	40.4%	9
Nipah	1998	513	398	77.6%	2
SARS	2002	8,096	774	9.6%	29
MERS*	2012	2,494	858	34.4%	28
COVID-19**	2020	222,642	9,115	4.1%	159

Sources: Johns Hopkins, CDC, World Health Organization, New England Journal of Medicine, Malaysian Journal of Pathology, CGTN
*As of November 2019 **As of March 19, 2020 at 7:30 am EST.
BUSINESS INSIDER

Figure 1: Surrounding text often highlights some information from the table but does not capture all data. Alternately, the linked text may be subjective or even misleading without the original table to check the claims.

Further, the structure of tables allows much more information to be presented in an efficient manner as humans can interpret meaning in the spatial relationship between cells. However, due to their challenging nature, recent algorithms have been less successful at extracting (Hoffswell and Liu, 2019) and understanding header and data structure in tables (Cafarella et al., 2018). In addition, any hierarchical and nested headers (common in printed documents) increases the difficulty in interpreting data cells, as shown in Figure 2.

In this paper, we propose to bridge this gap with statement verification and evidence finding using tables from scientific articles. This important task promotes proper interpretation of the surrounding article. In fact, the misunderstanding of tables can lead to the reporting of fake news that we see as being all too prevalent today.

* Equal Contribution
§ Corresponding Author

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 317–326
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

	n (% initiated smoking)	Unadjusted	
		OR (95% CI)	*p*
Baseline EC use			
Never	902 (8.2)	1.00	
Ever	21 (52.6)	**12.41 (4.53–33.99)**	**<.001**
Follow-up EC use			
No escalation	882 (8.1)	1.00	
Escalation	41 (41.0)	**7.94 (3.75–16.82)**	**<.001**
Age			
11–13	397 (4.4)	1.00	
14–15	270 (6.3)	1.45 (.71–2.97)	.312
16–18	256 (16.1)	**4.12 (2.19–7.76)**	**<.001**

Figure 2: A complex table sourced from (East et al., 2018) with hierarchical column and row structure. Additional difficulty follows from row hierarchy not being delineated by separate columns.

We present the first SemEval challenge to address table understanding. We introduce a brand new dataset of 1980 tables from scientific articles that addresses two challenging tasks important to table understanding:

A: Statement Fact Verification Given a statement, determine whether it is *supported*, *refuted* or *unknown* according to the table.

B: Cell Evidence Selection Given a statement, select the cells in the table that provide evidence supporting or refuting the statement.

The rest of this paper is formatted as follows: We first discuss related work. We then present a new large table understanding dataset containing close to 2000 tables that is the first to provide evidence labels at the cell level for statements and the first to focus on scientific articles. We provide a detailed analysis of the dataset including several baseline results. We then discuss the performance and approaches of the 19 participants in our challenge and end with an aggregated analysis of participating teams. Finally, we discuss future work.

2 Related Work

Natural Language Inference (NLI) The table evidence task can be best understood as a variation of the natural language inference task (Dagan et al., 2005), but on tabular data. NLI asks whether one (or more) sentence entails, refutes, or is unrelated to another sentence; our framing asks whether a given table entails, refutes, or is unrelated to a sentence. Several datasets have been created for studying NLI, such as SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), and SciTail (Khot et al., 2018).

Table QA This task is also closely related to the problem of search and question answering on tables. The closest example would be, given a table that is known to contain the relevant information, return cell values that answer a natural language question (Pasupat and Liang, 2015). A variation requires analyzing a collection of tables rather than a single one, along with the natural language question (Sun et al., 2016). Two of the most recent works are TAPAS (Herzig et al., 2020) and TaBERT (Yin et al., 2020), which jointly pre-train over textual and tabular data to facilitate table QA. However, such approaches have previously focused on traditional natural language questions ("What is the population of France?") rather than inference statements ("France has the highest population in Europe"), which may be entailed, refuted or unknowable from the given table.

Related Datasets The works closest to our dataset are TabFact (Wenhu Chen and Wang, 2020) and INFOTABS (Gupta et al., 2020). Both datasets were sourced from Wikipedia tables and contain hypothesis and premise pairs. TabFact has entailment and refute hypothesis types while INFOTABS has an additional "neutral" hypothesis category, much like our "unknown" statements. Both works show that neural models still lag far behind human performance for the fact checking task with tables.

While both datasets have been great at kindling interest in fact verification with tabular data, our dataset differs in two key aspects. First, we source from scientific articles in a variety of domains rather than Wikipedia infoboxes. Scientific tables have very specialized vocabulary and can be more difficult to interpret. Additionally, scientific tables have much more complex structure, like hierarchical column and row headers, rendering the assumption that the first column/row is the header unhelpful. Finally, tables are often directly referenced in scientific text unlike Wikipedia tables that are generally stand-alone. This creates an opportunity to leverage natural statements that depict the original author's style and intent. The second key differentiator of SEM-TAB-FACTS is the accompanying evidence annotations. We believe the future of fact verification and AI in general will be in cooperation with humans rather than in replacement. Thus, it is essential that models are able to present explanations for decisions on the relationship between the statement and table by showing the most relevant cells in a potentially very large table.

Source	#Tables	#Entailed	#Refuted	#Unknown	#Relevant	#Irrelevant
Train Crowdsourced	981	2,818	1,688	0	0	0
Train Auto-generated	1,980	92,136	87,209	0	1,039,058	15,467,957
Development	52	250	213	93	3,048	2,8495
Test	52	274	248	131	3,458	26,724

Table 1: Statistics for our SEM-TAB-FACTS dataset.

3 Dataset Details

Our dataset consists of two forms of generation: (1) a crowdsourced dataset, and (2) an auto-generated dataset. Table 1 presents the statistics of the dataset. We detail our dataset creation process in the following sections.

3.1 Data extraction and preprocessing

We sourced our tables from scientific articles belonging to active journals that are currently being published by Elsevier and are available on ScienceDirect[1]. We utilized Elsevier ScienceDirect APIs[2] to scrape scientific articles which belong to this list, and satisfy the following criteria: (1) the article is open-access[3], (2) the article is available under "Creative Commons Attribution 4.0 (CC-BY)" user license[4], and (3) the article has at least one table. We downloaded 1,920 articles belonging to 722 journals which contained 6,773 tables. We further filtered out complicated tables (e.g. multiple tables in a single table) using hand-written rules to get a set of 2,762 candidate tables from 1,085 articles for annotation. We also extracted sentences mentioning the table within the scientific article as candidate statements, which are corrected and then labeled manually by the annotators.

3.2 Crowdsourced labeling

The manually generated statements were collected using the crowdsourcing platform Appen[5]. We collected five entailed and five refuted statements for each table from the business preferred operators (BPO) on Appen. The BPO crowd is composed of employees hired by Appen on an hourly basis at a constant pay rate determined by Appen. We found that the workers were much more motivated for the task as they were able to ask questions if needed and we were also able to provide direct feedback to the workers. We initially attempted generating statements with workers from the Appen open-crowd, which is on-demand, but the quality was very poor as it was hard to automatically validate naturally generated statements. Our instructions explicitly lay out 7 types of statements and ask that workers attempt to make one of each type. We encourage the use of different sets of cells whenever possible. The types of statements are aggregation, superlative, count, comparative, unique, all and usage of caption or common sense knowledge. These are derived from the INFOTABS analysis (Gupta et al., 2020). We asked workers to avoid subjective adjectives like "best", "worst", "seldom" and look-up statements that only require reasoning with one cell. The pay for each statement set was 75 cents. In total, we collected 10000 statements for 1000 unique tables. See Figure 3 for an example table with its manually generated and natural statements.

Additionally, for our training data, we conducted a verification task to check for grammatical issues and doubly verify the statement label for both the generated and natural in-text statements. The verification task was paid at 3 cents per statement, which equates to 30 cents per table. We restricted the verification task to the workers in the open-crowd from English speaking countries. After verification, we only preserved the statements that were verified to be grammatically correct and the new label matched the original label. Natural statements were also verified in the same process. Although natural statements were generally factually correct, they were sometimes not able to be verified by the referenced table. Additionally, these statements often required rewording to ensure that all parts of the statement can be verified by the table, which was a step taken only for the development and test sets. This left us with 981 tables and 4506 statements. The majority of the removals were due to

[1] https://www.elsevier.com/__data/promis_misc/sd-content/journals/jnlactive.xlsx

[2] https://dev.elsevier.com/sd_apis.html

[3] https://www.elsevier.com/open-access

[4] https://www.elsevier.com/about/policies/open-access-licenses/user-licences

[5] https://appen.com/

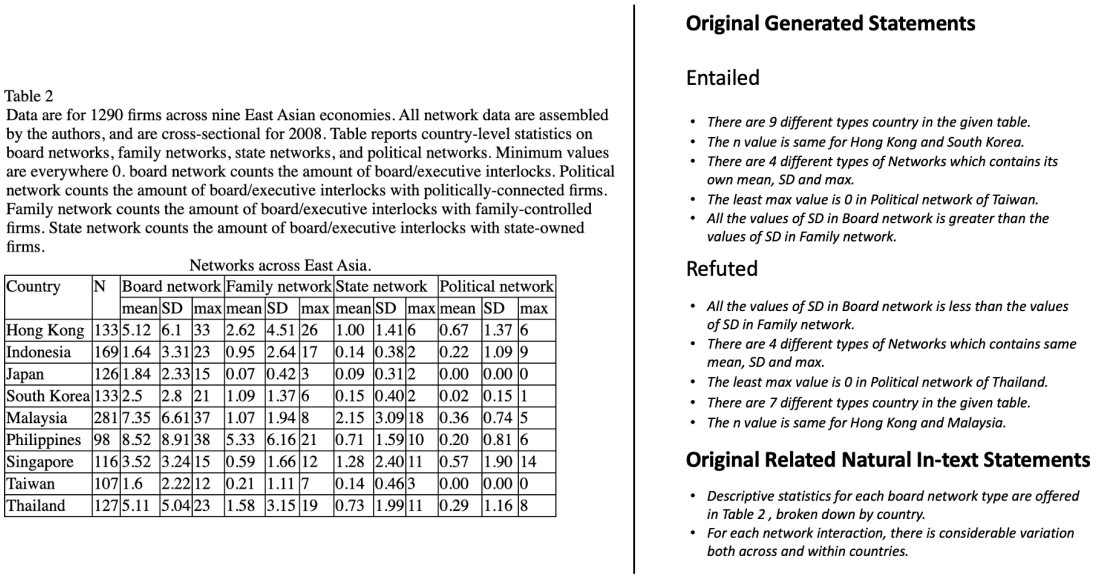

Table 2
Data are for 1290 firms across nine East Asian economies. All network data are assembled by the authors, and are cross-sectional for 2008. Table reports country-level statistics on board networks, family networks, state networks, and political networks. Minimum values are everywhere 0. board network counts the amount of board/executive interlocks. Political network counts the amount of board/executive interlocks with politically-connected firms. Family network counts the amount of board/executive interlocks with family-controlled firms. State network counts the amount of board/executive interlocks with state-owned firms.

Networks across East Asia.

Country	N	Board network			Family network			State network			Political network		
		mean	SD	max	mean	SD	max	mean	SD	max	mean	SD	max
Hong Kong	133	5.12	6.1	33	2.62	4.51	26	1.00	1.41	6	0.67	1.37	6
Indonesia	169	1.64	3.31	23	0.95	2.64	17	0.14	0.38	2	0.22	1.09	9
Japan	126	1.84	2.33	15	0.07	0.42	3	0.09	0.31	2	0.00	0.00	0
South Korea	133	2.5	2.8	21	1.09	1.37	6	0.15	0.40	2	0.02	0.15	1
Malaysia	281	7.35	6.61	37	1.07	1.94	8	2.15	3.09	18	0.36	0.74	5
Philippines	98	8.52	8.91	38	5.33	6.16	21	0.71	1.59	10	0.20	0.81	6
Singapore	116	3.52	3.24	15	0.59	1.66	12	1.28	2.40	11	0.57	1.90	14
Taiwan	107	1.6	2.22	12	0.21	1.11	7	0.14	0.46	3	0.00	0.00	0
Thailand	127	5.11	5.04	23	1.58	3.15	19	0.73	1.99	11	0.29	1.16	8

Figure 3: Sample crowd-sourced statements for one table (sourced from (Carney et al., 2020)). Please note that these are the original statements without any further corrections nor rephrasing.

grammatical errors as most BPO workers are not native English speakers. See Table 1 (first row) for detailed statistics of the crowd-sourced training set.

We initially attempted to collect the development and test sets as well as evidence annotations via the same method as the training set. However, we found that the quality was not gold-level and thus we (three of the authors) decided to manually correct the statements and annotate the evidence ourselves. All authors first annotated a small set of 102 statements to test inter-annotator agreement for statement relationship and evidence labeling. Out of 102 statements, we found 5 statements where at least one of three annotators disagreed on the relationship and a further 5 statements where the relationship was agreed but the evidence annotation differed. The other 92 were in complete agreement, indicating high agreement. Therefore, the annotations for the rest of the dev set were annotated by just one person. The test set was annotated fully by one author and the two other authors checked the annotations with all disagreements being resolved. See Figure 4 for a screenshot of the statement annotation correction and evidence annotation interface. See the third and fourth rows of Table 1 for detailed statistics of the dev and test sets.

3.3 Automatically generated statements

IBM Watson™ Discovery[6] is an AI-powered search and text analytics engine for extracting answers from complex business documents. One of the available functionalities is a Table Understanding service that produces a detailed enrichment of table data within an html document. We use this service to identify the body and header cells, as well as the *cell relationships*, within our dataset. We then proceed to use a set of templates to automatically create statements about each table. We begin by identifying which cells and columns are numeric and non-numeric using a simple regex. Unlike non-numeric cells, numeric cells and columns are appropriate for specific templates that expect numeric values, such as 'Value [V] is the maximum of Column [C]', where every value in column [C] has been identified as numeric. We also generate evidence for some of these templates. The template and evidence generation rules along with their inputs are detailed in Table 2. This process generated 3,512,978 statements from 1,980 tables which were highly skewed in favor of refuted statements. This dataset was then down-sampled to a maximum of 50 statements per table while ensuring a more even distribution between the two classes to form our final released auto-generated dataset. The

[6]https://www.ibm.com/cloud/watson-discovery

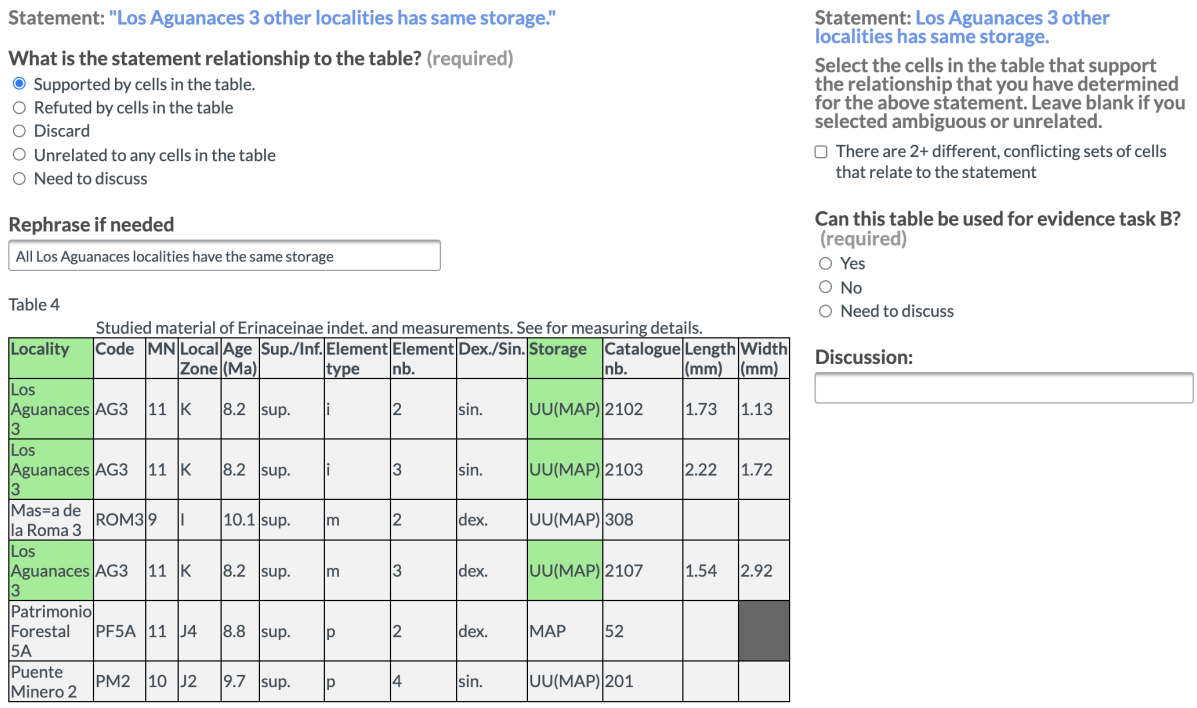

Figure 4: Screenshot showing the labeling interface for statement rephrasing, relationship labeling and evidence annotation.

full statistics for the auto-generated training data is shown in the second row of Table 1.

4 Evaluation Metrics

4.1 Task A: Statement Fact Verification

The goal of task A is to determine if a statement is entailed or refuted by the given table, or whether, as is in some cases, this cannot be determined from the table. We show two evaluation results. The first is a standard 3-way Precision / Recall / F1 micro evaluation of a multi-class classification that evaluates whether each table was classified correctly as Entailed / Refuted / Unknown. This tests whether the classification algorithm understands cases where there is insufficient information to make a determination. The second, simpler evaluation, uses the same P/R/F1 metric but is a 2-way classification that removes statements with the "unknown" ground truth label from the evaluation. The 2-way metric still penalizes misclassifying refuted/ entailed statement as unknown.

4.2 Task B: Cell Evidence Selection

In Task B, the goal is to determine for each cell and each statement, if the cell is within the minimum set of cells needed to provide evidence for the statement ("relevant") or not ("irrelevant"). In other words, if the table were shown with all other cells blurred out, would this be enough for a human to reasonably determine that the table entails or refutes the statement? The evaluation calculates the recall and precision for each cell, with "relevant" cells as the positive category. For some statements, there may be multiple minimal sets of cells that can be used to determine statement entailment or refusal. In such cases, our dataset contains all of these versions. We compare the prediction to each ground truth version and count the highest score.

5 Experiments

We present our baseline experimental setup for each task below.

Task A We employ state-of-the-art Table-BERT implementation[7] as proposed by Wenhu Chen and Wang (2020). We utilize Table-BERT's best performing configuration (Table-BERT-Horizontal-T+F-Template) as (1) using entity-linking to find the relevant columns for a statement, (2) flattening the table by scanning horizontally to form natural statements from the relevant columns and their cell values and (3) classifying the flattened table and the statement using the sentence pair classification

[7] https://github.com/wenhuchen/Table-Fact-Checking

Input	Template	Evidence	Example Statements
col_i, col_j	'The' + col_i_head + 'is' + col_i_val + ', when the' + col_j_head + 'is' + col_j_val	col_i_head, col_j_head, col_i_val, col_j_val	The Code is AG3 when the Locality is Los Aguanances3.
col	col_val + 'is in' + col_head	col_val, col_head for entailed; col for refuted	AG3 is in Code.
col	unique or same values	col for entailed; None for refuted	Sup./Inf. has the same values.
col[#]	'The maximum of' + col_head +'is'+val	col[#] for entailed; None for refuted	The maximum of Length(mm) is 2.22.
col[#]	'The minimum of' + col_head +'is'+val	col[#] for entailed; None for refuted	The minimum of Length(mm) is 1.54.
col[#]	'The mean of' + col_head + 'is' + val	col[#]	The mean of Length(mm) is 1.83.
col[#]	'The median of' + col_head + 'is' + val	col[#]	The median of Length(mm) is 1.73.
col[#]	'The mode of' + col_head + 'is' + val	col[#]	The mode of Length(mm) is 1.54, 1.73, 2.22.

Table 2: Template and evidence rules used for auto-generated ground truth. The examples are derived from Table 4 in Figure 4.

setting in BERT. To overcome the lack of unknown statements in our dataset, we supplement each table with randomly chosen statements from other tables. In Table-BERT, if the entity linking results in no matches, the flattened table is marked as [UNK]. As our dataset contains unknown statements, in such cases we consider all columns to be a match and flatten the entire table.

Using the above process, we perform the following experiments (1) apply the Table-BERT model out-of-the-box (2) re-train Table-BERT model with unknown statement and apply on our test data (3) fine-tune the model in (2) with our manual+auto-generated data and apply on our test data. We also compare these experiments with a majority baseline with entailed as our majority class. The results are presented in Table 3. Applying Table-BERT model out-of-the-box provides some improvement over a majority-baseline. However, when the model is retrained with previously missing unknown statements, the performance improves for three-way classification. Further fine-tuning the model with our training dataset (both manual and auto-generated) provides the best performance on the two-way F1-score.

Task B We present the following two baselines for Task B: (1) a random baseline where each cell is marked relevant or irrelevant randomly (2) a simple word-match-based baseline where a cell is marked relevant if it overlaps with the statement. The baseline results are presented in Table 4.

Experiment	Test	
	2-way	3-way
majority-baseline	52.42	42.16
original Table-BERT	56.77	45.58
re-trained Table-BERT	52.96	48.33
+ FT with SEM-TAB-FACTS	56.81	48.24

Table 3: Task A baseline results using F1-score.

Experiment	Dev	Test
random-baseline	21.18	20.47
word-match	49.53	47.39

Table 4: Task B baseline results using F1-Score

6 Competition Results

We present two leaderboards for each task[8]. The official leaderboard is from participants who have given us detailed descriptions on their system and affirmed that they did not incorporate any information from the test set that changed their final model. This is a more accurate representation of system quality. The unverified leaderboard is composed of participants who either did not give enough detail or have affirmed that they incorporated some test data information in their final model. The participants did not have access to labels for test data but some teams altered their models upon examining

[8]We made the assumption that teams would not make any use of the test data, as is usually the case for algorithm evaluation, but we did not make this explicit ahead of time and some teams did not realize this was an issue. We decided to have two leaderboards to have a fair comparison for all teams.

Team	3-way F-Score	2-way F-Score
Official Leaderboard		
King001	**84.48**	**88.74**
THiFly_Queen	83.76	84.55
RyanStark	81.51	87.22
sattiy	77.32	84.96
BreakingBERT@IITK	69.31	76.81
Volta	67.34	72.89
TAPAS	66.81	73.13
AttesTable	65.59	71.72
Yaoxu	60.76	75.8
Beary-group	58.37	72.56
ok-team	57.79	71.84
SUNLP	47.92	59.58
FishToucher	41.83	52.01
KaushikAcharya	36.23	23.08
Unverified Leaderboard		
Skywalker	92.55	95.15
MagicPai	90.88	94.03
endworld	82.35	88.16
Paima	81.96	88.85
ravikranc	57.90	71.99

Table 5: Task A Leaderboard

Team	F-Score
Official Leaderboard	
BreakingBERT@IITK	**65.17**
Volta	62.95
King001	62.14
FishToucher	60.06
RyanStark	54.96
Sattiy	48.56
AttesTable	43.02
KaushikAcharya	33.81
Unverified Leaderboard	
MagicPai	88.74
SkyWalker	73.05
endworld	57.85
Paima	51.97

Table 6: Task B Leaderboard

the input data in the test set. Although we discouraged this approach, we present the results in hopes it can give some interesting information about how much improvement might be possible with having access to input test data.

19 teams participated in Task A. Of the 14 teams on the official leaderboard, King001 obtained the highest score for task A for both the 2-way (88.74) and 3-way (84.48) F-scores. However, the top three participants have comparable scores. All teams except for the last two beat our best baseline in Table 3. The unverified leaderboard includes 5 teams and contains higher scores thank in the official leaderboard. However, due to the reasons outlined above, we cannot say with certainty that the results are reproducible. The full leaderboard results for all participants are in Table 5.

Task B is a much harder task and fewer teams participated in this challenge. Of the 12 teams that participated, 8 are in the official leaderboard. The best score is 65.17 by BreakingBERT@IITK(65.17) which is noticeably lower than the F-scores in Task A. Similarly to Task A the results in the unverified leaderboard are considerably higher. The full leaderboard results for all participants are in Table 6.

We summarize the system details for all participating teams in Tables 7 (Task A) and 8 (Task B). In general, deep learning was the most popular approach used by the participants e.g. BiLSTM with attention, BERT (Devlin et al., 2019) etc. Most of the participants used transformer-based models to train their systems with flavors ranging from general-domain BERT (Devlin et al., 2019) to table-understanding specific versions like TAPAS (Herzig et al., 2020), TaBERT (Yin et al., 2020) and Table-BERT (Wenhu Chen and Wang, 2020). One third of the participants employed some form of ensembling technique in their submission.

Most of the participants have used the manually generated ground-truth in the development of their systems, with only one team not finding it useful. Further, a large percentage of participants have used the auto-generated ground truth in their systems with three teams not finding it helpful in their evaluation.

In terms of external resources, a majority of the participants used external table understanding resources in their systems. Further, most of the participants employed pre-processing techniques like acronym completion, removing special characters, etc... A substantial percentage of participants used techniques like incorporating word embeddings, entity resolution etc. Finally, a large number of participants used TabFact (Wenhu Chen and Wang, 2020) as an external dataset.

We also conducted additional analyses on participant submissions on the official leaderboard. We show through the average confusion matrix for Task A in Table 9 that the Unknown label was the most difficult. In fact, there were more unknown statements incorrectly labelled as entailed than were correctly categorized. Naturally, the statements with the lowest accuracy ($< 25\%$) consist of mainly unknown statements, especially those statements

Team	Description
AttesTable (Varma et al., 2021)	Extended TAPAS to 3 classes by fine-tuning it. Employed a novel way of synthesizing "unknown" samples.
BreakingBERT@IITK (Jindal et al., 2021)	Ensemble models with TAPAS and TableBERT Transformers in a hierarchical two-step method for 3-way classification (unknown vs not unknown first)
Beary-group	Used TAPAS model with TabFact task, and added unique features. Employed prepossessing tricks like k-fold validation and replacing the characters and did hyperparameter tuning.
BOUN (Köksal et al., 2021)*	Used text augmentation techniques such as back translation and synonym swapping on the TAPAS model. Domain adaptation and joint learning using SemTabFacts and TabFact datasets.
endworld	Data Cleaning. Ensemble combining 80 instances of trained TaPas-Large and label smoothing.
FishToucher	Motivated by TaPas, used BERT and enriched the embedding layer with two new token type embeddings: row and column ids* (*The team mistakenly submitted an old model version, see paper for more accurate scores)
Kaushik Acharya (Acharya, 2021)	Parsed statements into candidate logical form; mapped result to handwritten rules, to then execute over relevant cells (identified using string matching and universal dependency parsing)
King001	Trained 20 instances of TaPas, SAT and Table-Bert for an ensemble of 60 models. Used preprocessing like acronym completion, rules to align the table content with the question content, label smoothing.
MagicPai	Multi-model training using models such as TaBERT, tapas_wikisql, tapas_TabFact, tapas_masklm. Finally rule amendments and aligning the distribution of training and test data
ok-team	TAPAS pretrained on TabFact with preprocessing of data (like transforming English numerals to Arabic numerals, removing special characters etc.)
Paima	Fine-tuned TAPAS optimized to perform window scanning on statement-related table data. Preprocessing to reduce abbreviations for table headers, and identifying operation expressions.
RyanStark	Multi-model TaBERT pretrained Model fusion. Pre-processing such as case and abbreviations.
Sattiy (Ruan et al., 2021)	Ensemble of 6 fine-tuned pre-trained models on the augmented data with content snap-shot input. Augmented the data provided by expanding the labels. Used Fast Gradient Method and added disturbance to the embedding layer to obtain a more stable word representation and a more general model.
SkyWalker	Deep learning, LPA rules, TAPAS dataset
SUNLP	BERT for sequence classification, transfer learning
TAPAS (Müller et al., 2021)	Ensemble of TAPAS (BERT-large-like) models: trained with a Mask-LM task on Wikipedia tables, intermediate pre-training data and TabFact data. Hierarchical two-step method for 3-way classification. Added neutral statements during training: random and by removing one of the evidence columns.
THiFly_Queen (Yuxuan et al., 2021)	Ensemble models in a hierarchical two-step method. 8-model to identify unknown statements and 9-model ensemble to classify entailed/refuted. Incorporated different ensemble weights for various statement types (count, superlative, unique).
Volta (Gautam et al., 2021)	Finetuned TAPAS that was pretrained on TabFact. Pre-processing to standardize multiple header rows to a single header.
Yaoxu	Added numeric and enumerate features to TAPAS and also statistic information (such as count) as a new row/column to the table.

Table 7: Descriptions of systems from participants for Task A. *Note: Team BOUN did not participate in the official leaderboard.

Team	Description
BreakingBERT @IITK	An ensemble of an individual cell-based NLI approach and a similarity approach with the cells and statement
FishToucher	BERT CLS tokens for statement and table cells are used to determine cell relationships to each other, and the statement (for relevant cells)
Kaushik Acharya	Relevant cells are output as part of Task A
RyanStark	BOW approach with rules applied based on word matches in header and data cells.
Volta	Finetuned TAPAS for cell selection. Different models for entailed and refuted statements. Used transfer learning and header standardization.

Table 8: Descriptions of systems from participants for Task B (when provided)

	Refuted	Entailed	Unknown
Refuted	164	81	3
Entailed	46	226	2
Unknown	16	72	43

Table 9: Task A average confusion matrix

that have words overlapping with those in the table. Out of the entailed and refuted statements, ones that require numerical reasoning, like range, count or comparisons seemed to be most challenging. The statements with the highest accuracy ($> 95\%$) generally had most words or numbers exactly overlapping with those in the table. In task B, out of the statements with less than 30% evidence F-score, 86% were ones with a refuted relationship. Conversely, the statements with greater than 70% F-score, 74% were ones with an entailed relationship. This shows that it is more difficult to find the most direct evidence to prove that a statement is refuted by a table than it is to show the positive evidence that a particular statement is supported by it. We believe this is an interesting line of research for future studies.

7 Conclusion and Future Works

In this paper, we presented the data and competition results for SEM-TAB-FACTS, Shared Task 9 of SemEval 2021. We created a large dataset via automated and crowdsourced fact verification as well as evidence finding for tables. Our 19 teams had a variety of techniques to tackle this unique but very relevant problem. The evidence finding scores are still quite low and have a large improvement potential. Additionally, the test set may be expanded in future versions of this task with a combination of manually generated, natural, and automated statements.

References

Kaushik Acharya. 2021. Kaushikacharya at SemEval-2021 task 9: Candidate generation for fact verification over tables. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.

Ruobing Su Aylin Woodward and Shayanne Gal. 2020. What to know about the coronavirus outbreak in 17 charts and maps. *Business Insider*.

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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Michael Cafarella, Alon Halevy, Hongrae Lee, Jayant Madhavan, Cong Yu, Daisy Zhe Wang, and Eugene Wu. 2018. Ten years of webtables. *Proceedings of the VLDB Endowment*, 11(12):2140–2149.

Richard W Carney, Travers Barclay Child, and Xiang Li. 2020. Board connections and crisis performance: Family, state, and political networks. *Journal of Corporate Finance*, 64:101630.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katherine East, Sara C Hitchman, Ioannis Bakolis, Sarah Williams, Hazel Cheeseman, Deborah Arnott, and Ann McNeill. 2018. The association between smoking and electronic cigarette use in a cohort of young people. *Journal of Adolescent Health*, 62(5):539–547.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021. Volta at SemEval-2021 task 9: Statement verification and evidence finding with tables using TAPAS and transfer learning. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

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.

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.

Jane Hoffswell and Zhicheng Liu. 2019. Interactive repair of tables extracted from pdf documents on mobile devices. In *ACM Human Factors in Computing Systems (CHI)*.

Aditya Jindal, Ankur Gupta, Jaya Srivastava, Preeti Menghwani, Vijit Malik, Vishesh Kaushik, and Ashutosh Modi. 2021. Breakingbert@iitk at SemEval-2021 task 9: Statement verification and evidence finding with tables. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI Conference on Artificial Intelligence*.

Abdullatif Köksal, Yusuf Yüksel, Bekir Yıldırım, and Arzucan Özgür. 2021. BOUN at SemEval-2021 Task 9: Text Augmentation Techniques for Fact Verification in Tabular Data. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Thomas Müller, Julian Martin Eisenschlos, and Syrine Krichene. 2021. TAPAS at SemEval-2021 task 9: Reasoning over tables with intermediate pretraining. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*. International Committee 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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Xiaoyi Ruan, mei Jin, Jian Ma, Lianxin Jiang, Mo Yang, and Jianping Shen. 2021. Sattiy at SemEval-2021 task 9: Method for statement verification and evidence finding with tables based on multi-model ensemble. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.

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*, WWW '16, page 771–782, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Harshit Varma, Aadish Jain, Pratik Ratadiya, and Abhishek Rathi. 2021. Attestable at SemEval-2021 task 9: Extending statement verification with tables for unknown class, and semantic evidence finding. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.

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

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.

Zhou Yuxuan, Zhou Kaiyin, Liu Xien, Wu Ji, and Zhu Xiaodan. 2021. Thifly_queen at SemEval-2021 task 9: Statement verification and evidence finding with tables. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*.