SemEval 2021 (Task 9): Statement Verification and Evidence Finding with Tables

CS779-A Project

Group 5:

Vijit Malik (170791)- vijitvm@iitk.ac.in

Vishesh Kaushik (170805)- kvishesh@iitk.ac.in

Aditya Jindal (170048)- adityaji@iitk.ac.in

Problem Statement

The task will have two subtasks to explore table understanding:

Subtask A : **Table Statement Support**

- Does the table support the given statement?

Subtask B : **Relevant Cell Selection**

- Which cells in the table provide evidence for the statement?

Task Link: https://sites.google.com/view/sem-tab-facts

Codalab Link: https://competitions.codalab.org/competitions/27748

Subtask A : Table Statement Support

Given a statement and a table, determine whether the statement is supported by the table.

In this classification problem, a statement is assigned one of the following labels:

1. **Fully Supported:** Statement is supported by data found within the table.
2. **Refuted:** Statement is contradicted by table.
3. **Unknown:** Not enough information in table to assess statement veracity.

Subtask B: Relevant Cell Selection

Given a statement and a table, find which table cells form relevant evidence for the statement (if any).

A table cell is evidence for a statement if it helps support or refute a part of the statement:

1. **Relevant**: the cell must be included.
2. **Ambiguous:** the cell is allowed to be either included or not included.
3. **Irrelevant:** the cell must not be included.

Sample SemEval XML File

Table ID

Caption/Table Metadata

Cell Information

```
<table_1>
 <caption text="
Overview stimuli EAST.
">
 </caption>
 <row row="0">
  <cell col="0" row="0" text="Bodily sensations">
   <evidence statement_id="0" type="relevant" version="0">
   </evidence>
   <evidence statement_id="1" type="irrelevant" version:="0">
   </evidence>
   <evidence statement_id="2" type="irrelevant" version:="0">
   </evidence>
  </cell>
  <cell col="1" row="0" text="Agoraphobic situation">
   <evidence statement_id="0" type="irrelevant" version:="0">
   </evidence>
   <evidence statement_id="1" type="irrelevant" version:="0">
   </evidence>
   <evidence statement_id="2" type="irrelevant" version:="0">
   </evidence>
  </cell>
```

Task 2: For Statement with ID 1, indicate whether this cell is related to whether the table entails or refutes the statement. The ground truth may have multiple versions but only one version should be provided from participants.

Term	Type	Factor	Example
Extremely	Intensifier	2.0	Made me settle very quickly. *Extremely* reliable application
			Extremely slow loading and shows database connection error.
Absolutely	Intensifier	1.75	*Absolutely* useless app. The app is useless, never works properly.
			An *absolutely* useful and friendly application!
Quite	Downtoner	0.75	*Quite* useful app for Dubai residents.
			It's *quite* frustrating as I am unable to use the app
			Great interface. The app is *quite* slow but has good functionalities.
Pretty	Downtoner	0.50	*Pretty* good but some bugs
			Poorly designed app. *Pretty* much useless and probably just a media stunt.
Always	Intensifier	1.5	App doesn't work. It *always* shows error in login.
			New version *Always* crashes.

Subtask-B Output: A sample table that shows the correct results for subtask B where : green = relevant, red = irrelevant, purple = ambiguous

Subtask-A output

Statement	Label
The polarity score of the opinion word that follows the downtoner "quite" is multiplied by the factor (0.75)	Supported
"New version Always crashes" is an example for "Quite"	Refuted
The "Extremely" term has the highest factor.	Supported
The polarity score of the opinion word that follows the intensifier "very" is multiplied by the factor (1.25)	Unknown

Why this task is worth doing ?

- Tables are ubiquitous in documents and presentations for conveying important information in a concise manner.
- The misunderstanding of tables can lead to report fake news.

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

Linguistic vs Symbolic Reasoning

- These two aspects of Fact checking in tabular data differ significantly.
- Linguistic Reasoning: Requires more of semantic-level understanding of text.
- Symbolic Reasoning: Requires symbolic execution on the table structure.

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

Source: Chen, Wenhu, et al. "TabFact: A large-scale dataset for table-based fact verification." *arXiv preprint arXiv:1909.02164* (2019).

Related Work

S.No.	Paper	Conference
1.	TabFact:A Large-scale Dataset for Table-based Fact Verification	ICLR 2020
2.	LogicalFactChecker: Leveraging Logical Operations for Fact Checking with Graph Module Network(SOTA)	ACL2020
3.	TAPAS: Weakly Supervised Table Parsing via Pre-training	ACL 2020
4.	Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision	ACL 2017

TabFact : A Large-scale Dataset for Table-based Fact Verification

Approaches Used

1. **Table-BERT**: Encodes table and statements into a linearized input similar to NLI.
2. **Latent Program Algorithm (LPA)**: Statements are semantically parsed against tables into a program type format.

Source: Chen, Wenhu, et al. "TabFact: A large-scale dataset for table-based fact verification." *arXiv preprint arXiv:1909.02164* (2019).

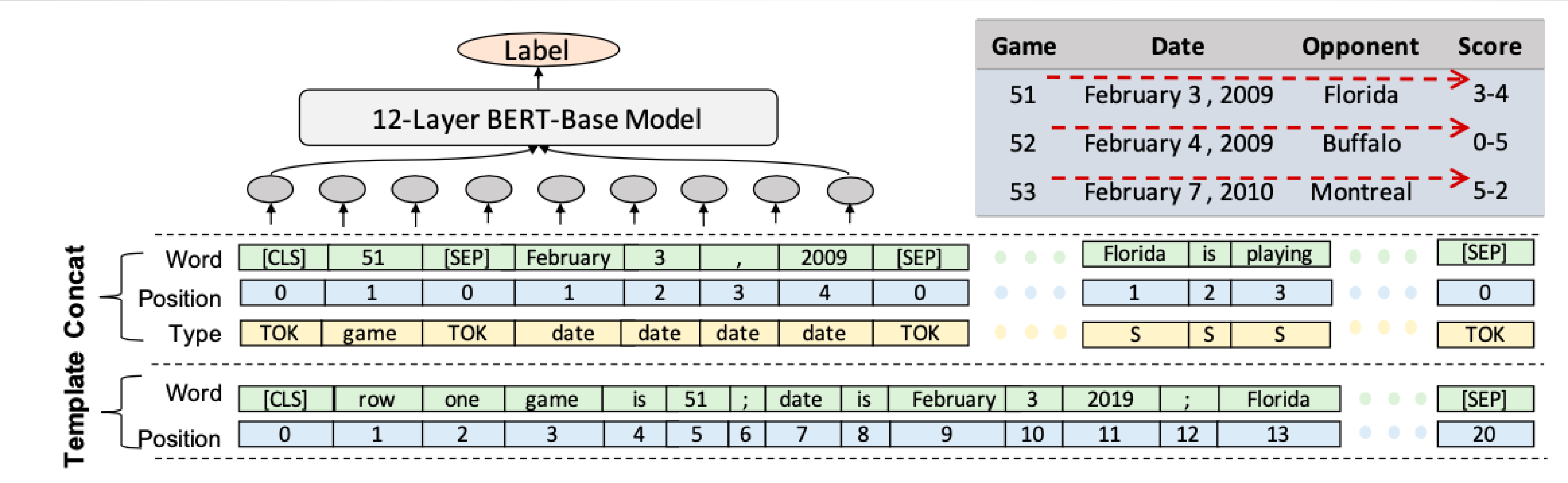
Table-BERT

Encoding table and statements as premise and hypothesis like in NLI tasks.

Important features:

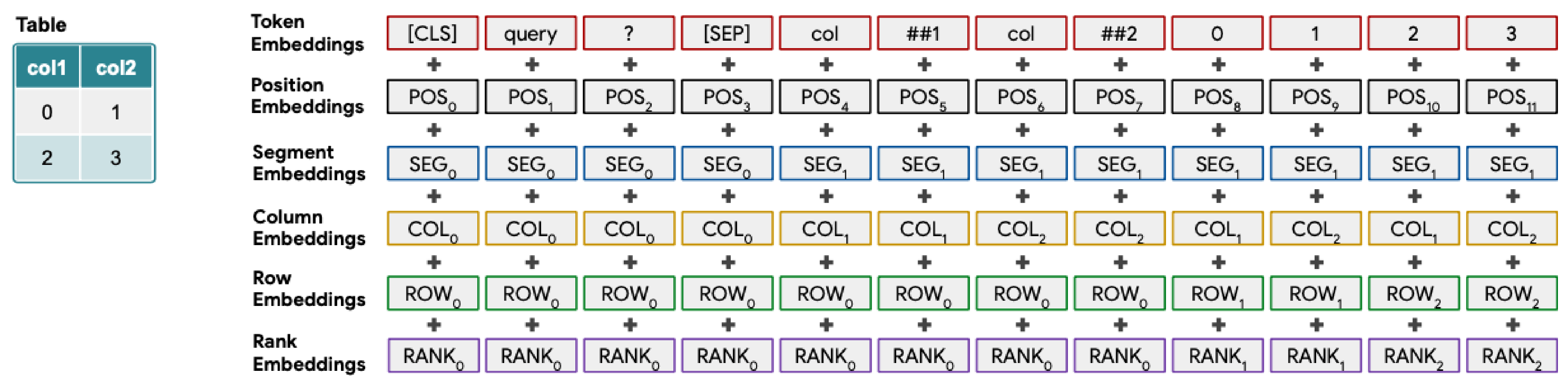
1. Shrinking the linearized table using ***entity linking.***
2. Two linearization methods: Concat and Template

Template vs Concat Linearization in Table-BERT



TAPAS: Weakly Supervised Table Parsing via Pre-training

Pre-training Structure



Source: Herzig, Jonathan, et al. "TAPAS: Weakly Supervised Table Parsing via Pre-training." *arXiv preprint arXiv:2004.02349* (2020).

Fact Checking vs Table QA

- Question itself provides strong signals needed for answer type and span identification.
- The fact or statement provided is false even if some part of it is wrong.
- The facts can be conjuctive due to which they need to be broken down and verified individually.

Results

Model	Val	Test	Test (simple)	Test (complex)	Small Test
Human Performance	-	-	-	-	92.1
Majority Guess	50.7	50.4	50.8	50.0	50.3
BERT classifier w/o Table	50.9	50.5	51.0	50.1	50.4
Table-BERT (Horizontal-S+T-Concatenate)	50.7	50.4	50.8	50.0	50.3
Table-BERT (Vertical-S+T-Template)	56.7	56.2	59.8	55.0	56.2
Table-BERT (Vertical-T+S-Template)	56.7	57.0	60.6	54.3	55.5
Table-BERT (Horizontal-S+T-Template)	66.0	65.1	79.0	58.1	67.9
Table-BERT (Horizontal-T+S-Template)	66.1	65.1	79.1	58.2	68.1
LPA-Voting w/o Discriminator	57.7	58.2	68.5	53.2	61.5
LPA-Weighted-Voting w/ Discriminator	62.5	63.1	74.6	57.3	66.8
LPA-Ranking w/ Discriminator	65.2	65.0	78.4	58.5	68.6
LogicalFactChecker (program from LPA)	71.7	71.6	**85.5**	64.8	74.2
LogicalFactChecker (program from Seq2Action)	**71.8**	**71.7**	85.4	**65.1**	**74.3**

SemEval-Dataset Description

- Data is sourced from open access scientific articles with tables using APIs provided by Science Direct .
- The statements sourced from automatic generation, the surrounding article text and crowdsourcing.
- Each statement adapted from existing text and verified by at least one reader.
- The format that the data will be procured is in XML so that the tables will be structured.

TabFact vs SemEval data

- Fewer tables in SemEval Data (3k tables vs. TabFact's 16k tables).
- TabFact data sourced from wikipedia tables, while SemEval data sourced from scientific articles .
- Subtask B requires models to show evidence for prediction.
- Number of statements in SemEval is 185k and TabFact has 118k statements.

SemEval-Dataset Statistics

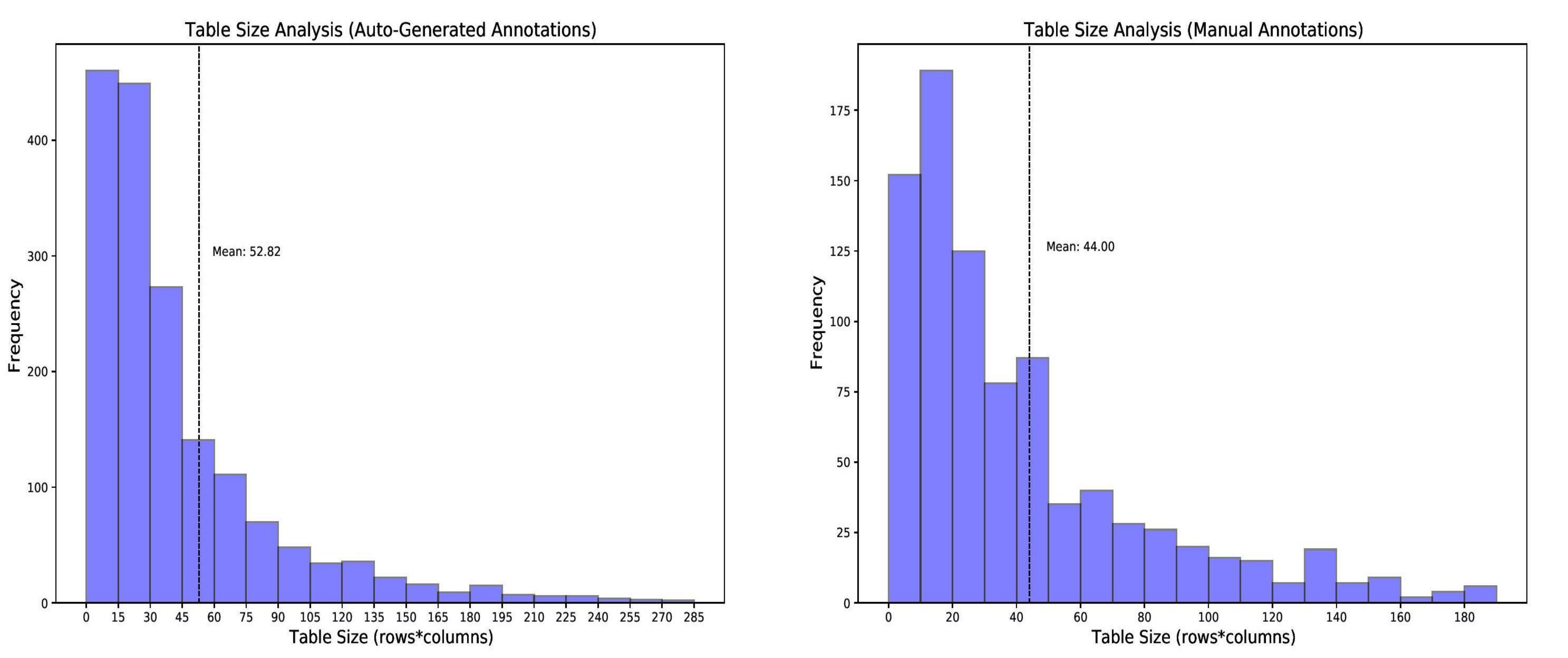
Dataset Type (Provided by TA)	Number of Tables			Mean No. Of Rows	Mean No. Of Columns	Mean No. Of Cells	Mean No. Of Statements
	Train	Dev	Test				
Auto-Generated	1591	195	194	10.23	4.79	52.82	90.58
Manual	783	100	98	9.23	4.65	44	4.59

Dataset Type (Original Semeval V1.3)	Number ofTables	
	Train	Dev
Auto-Generated	1980	52
Manual	981	

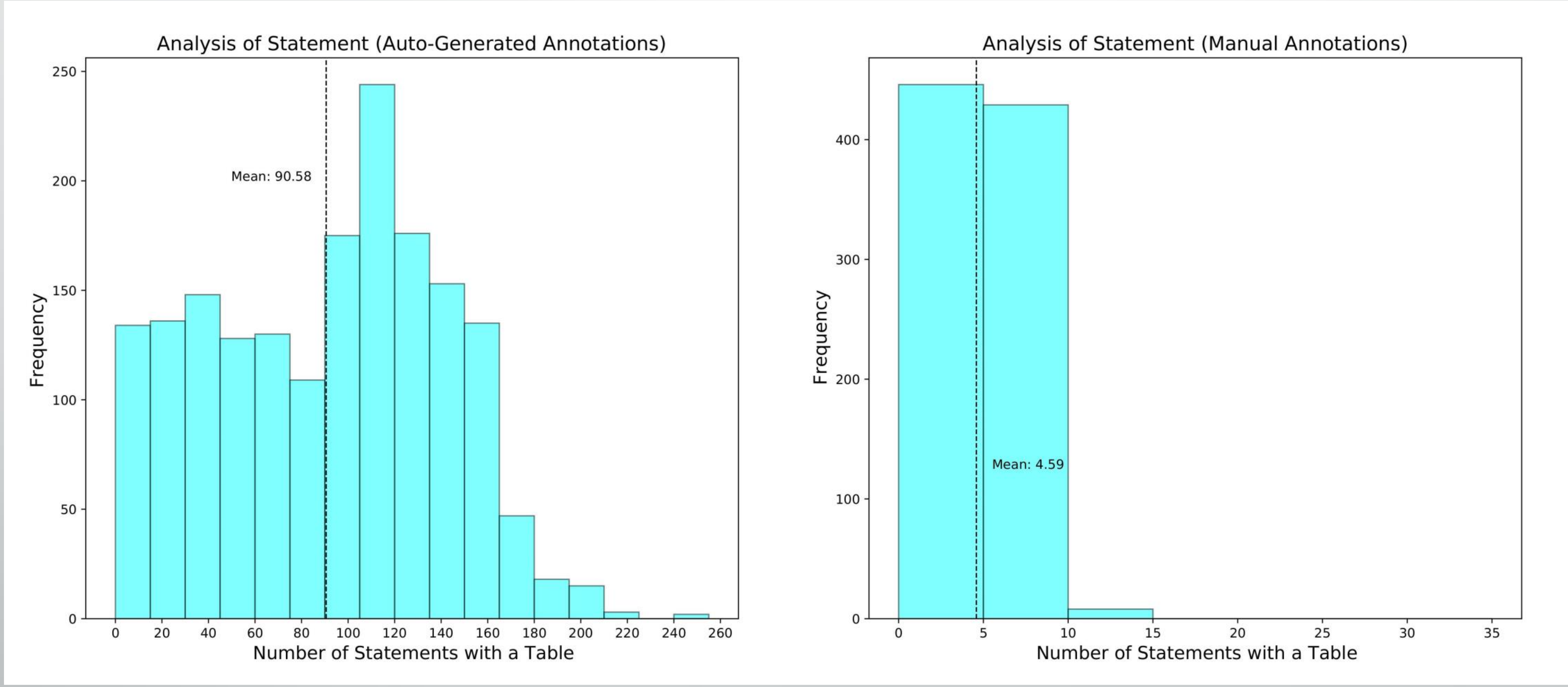
Manual vs Autogenerated Statements

- Statements are divided into autogenerated and manual , where manual statements are relatively difficult to classify.
- There are 1980 tables in Autogenerated dataset 981 tables in Manual dataset
- Manual statements are difficult to generate and hence account for ~ 4510 statements whereas autogenerated account for a large 179300 statements .

Table Size statistics of Manual vs Auto generated statements dataset



Number of Statements for a table in Manual vs Auto generated statements dataset



Evalution Metrics

- **Subtask A**
  - Simpler evaluation will remove statements with the "unknown" ground truth label .
  - Metric will still penalize misclassifying Refuted/ Entailed statement as unknown. The score used for ranking is the F1 score.
- **Subtask B**
  - F1 score for each cell, with "relevant" cells as the positive category.
  - The score will be averaged over all statements in each table first, before proceeding to average across all tables.

Methodology for subtask A

- Evaluated using Pre-trained TAPAS model fine-tuned over TabFact dataset.
- Implemented TableBERT and other table transformers based on SciBERT and RoBERTa.
- Implemented BiGRU layers on top of TableRoBERTa.

Results for Subtask A (Dev Set)

Model	Train set	Dev Set	Metrics (On Dev Set)			
			Precision	recall	F1	Acc(%)
TableBERT	**Auto**	**Auto**	**0.875**	**0.859**	**0.867**	**86.00**
TableRoBERTa	Auto	Auto	0.635	0.631	0.633	64.05
TableRoBERTa+BiGRU	Auto	Auto	0.647	0.634	0.660	67.13
TableSciBERT	Auto	Auto	0.710	0.698	0.643	65.64
TAPAS	TabFact	Auto	0.667	0.620	0.732	74.76
TableBERT	Auto	Manual	0.588	0.582	0.585	58.95
TableRoBERTA	Auto	Manual	0.529	0.507	0.518	51.95
TableRoBERTa+BiGRU	Manual	Manual	0.542	0.520	0531	53.28
TableSciBERT	Manual	Manual	0.614	0.606	0.610	61.07
TAPAS	**TabFact**	**Manual**	**0.778**	**0.761**	**0.771**	**72.26**

Test Results For Subtask A

Model	Test Dataset	Precision	Recall	F1	Accuracy(%)
TableBERT	Auto	0.8717	0.8718	0.8717	87.17
Group 15 (TAPAS)	Auto	0.9696	0.9559	0.9627	96.23
TAPAS	Manual	0.8264	0.7157	0.7671	70.83
Group 15 (TAPAS+ Table bert+ scibert)	Manual	0.7975	0.8497	0.8227	75.44

Methodology for subtask B

- Done as an individual cell based Natural Language Inference task.
- The premise is taken as the combination of row header, column header and cell contents.
- Hypothesis is taken as the statement provided.

[CLS] + [Row_Header] + [Cell Content] + [Column Header] + **[SEP]** + [Statement Text] + **[SEP]**

Results for Subtask B

Model	Train Set	Test Set	F1 score
CellBERT	Auto	Auto	0.7047
Group 15	Auto	Auto	0.5789

Model	Train Set (Auto)	Dev Set (Auto)
	F1	F1
CellBERT	0.7772	0.6783

Result Analysis

- TableSciBERT was underperforming compared to TableBERT on Subtask A even when the dataset had a lot of scientific statements.
- F1 score on TAPAS is greater than the accuracy for Subtask A suggesting that we've got a good amount of sensitivity for a class.
- We used less amount of data to train CellBERT and even that provides us with promising results.

Challenges in Subtask A

- Simple NLI based model failed to capture the logical connections between the cells.
- All existing works based on 2 labels rather than 3, and so difficult to use pretrain model.
- Lack of computational resources makes it difficult to use models like LPA .

Challenges in Subtask B

- No prior work has been done on subtask B.
- 3 classes for each level of relevance: Relevant, Ambiguous and Irrelevant.
- Used a considerably less amount of data to train CellBERT.

Conclusion

- Try to look for a solution to an under-explored but important problem: Statement Verification and Evidence Finding with Tables.
- We verified the existing models like TableBERT and TAPAS.
- We implemented TableSciBERT and TableRoBERTa by putting Bi-GRU layers on top of it.

Future Work

- In the future, we plan to implement new models that can tackle both linguistic and symbolic reasoning.
- Fine tune TAPAS on the manual and auto generated dataset
- Use an ensemble to achieve a boost in accuracy for subtask A.
- In case of subtask B, we like to experiment on other NLI techniques and models.
- Use more data for training CellBERT.

Thank You