An Overview of Tabular Data Extraction

Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context

Xinyi Zheng; University of Michigan
Douglas Burdick, Lucian Popa; IBM Research-Almaden
Xu Zhong; IBM Research Australia
Nancy Xin, Ru Wang; IBM Research-Almaden

Motivation

- 2.5 - 4 trillion documents in PDF format!
- Tables - best way to summarize, aggregate, compare data.
- Tabular data in the PDF is helpful in data analytics, data-driven decisions.
- Significant time and manual efforts to extract data.
- Importing the data in relational database facilitates answering user queries.

WHY “HARNESSING” !

- Table Extraction
  - Input: PDF page / Image / Office Documents
  - Output:
    - Table border for each table
    - Partitioning table contents into cells
    - Both vertical and horizontal alignment of cells
- Table Understanding
  - Input:
    - Output from Table Extraction Module (Document with Explicit Table Structure)
  - Output:
    - Annotate cells and tables with semantic info pertaining to them in a form amenable to post-processing

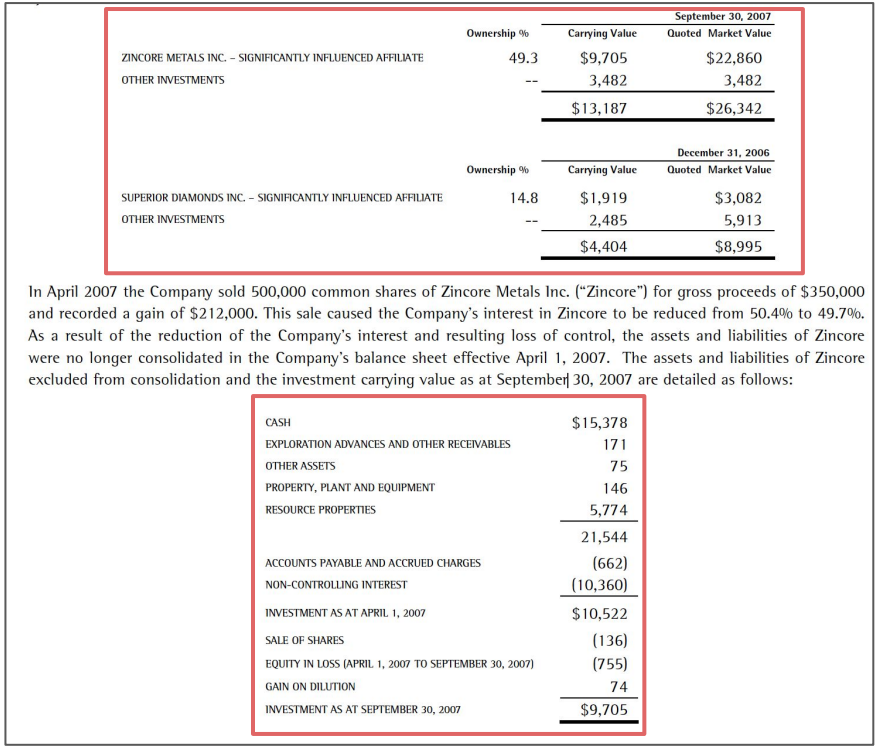
Challenges

- Different table types.
  - Matrix table - values have both row and column headers.
  - List/Entity table - values have only row headers.
  - Relational table - values have only column headers.

Challenges

- Different table types.
  - Matrix table - values have both row and column headers.
  - List/Entity table - values have only row headers.
  - Relational table - values have only column headers.

- Multiple tables and table types on same page.



	Ownership %	September 30, 2007 Carrying Value	Quoted Market Value
ZINCORE METALS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	49.3	$9,705	$22,860
OTHER INVESTMENTS	--	3,482	3,482
		$13,187	$26,342

	Ownership %	December 31, 2006 Carrying Value	Quoted Market Value
SUPERIOR DIAMONDS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	14.8	$1,919	$3,082
OTHER INVESTMENTS	--	2,485	5,913
		$4,404	$8,995

In April 2007 the Company sold 500,000 common shares of Zincore Metals Inc. ("Zincore") for gross proceeds of $350,000 and recorded a gain of $212,000. This sale caused the Company's interest in Zincore to be reduced from 50.4% to 49.7%. As a result of the reduction of the Company's interest and resulting loss of control, the assets and liabilities of Zincore were no longer consolidated in the Company's balance sheet effective April 1, 2007. The assets and liabilities of Zincore excluded from consolidation and the investment carrying value as at September 30, 2007 are detailed as follows:

CASH	$15,378
EXPLORATION ADVANCES AND OTHER RECEIVABLES	171
OTHER ASSETS	75
PROPERTY, PLANT AND EQUIPMENT	146
RESOURCE PROPERTIES	5,774
	21,544
ACCOUNTS PAYABLE AND ACCRUED CHARGES	(662)
NON-CONTROLLING INTEREST	(10,360)
INVESTMENT AS AT APRIL 1, 2007	$10,522
SALE OF SHARES	(136)
EQUITY IN LOSS (APRIL 1, 2007 TO SEPTEMBER 30, 2007)	(755)
GAIN ON DILUTION	74
INVESTMENT AS AT SEPTEMBER 30, 2007	$9,705

Challenges

- Presence of “other” data between tables.

	Ownership %	September 30, 2007 Carrying Value	Quoted Market Value
ZINCORE METALS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	49.3	$9,705	$22,860
OTHER INVESTMENTS	--	3,482	3,482
		$13,187	$26,342

	Ownership %	December 31, 2006 Carrying Value	Quoted Market Value
SUPERIOR DIAMONDS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	14.8	$1,919	$3,082
OTHER INVESTMENTS	--	2,485	5,913
		$4,404	$8,995

In April 2007 the Company sold 500,000 common shares of Zincore Metals Inc. (“Zincore”) for gross proceeds of $350,000 and recorded a gain of $212,000. This sale caused the Company’s interest in Zincore to be reduced from 50.4% to 49.7%. As a result of the reduction of the Company’s interest and resulting loss of control, the assets and liabilities of Zincore were no longer consolidated in the Company’s balance sheet effective April 1, 2007. The assets and liabilities of Zincore excluded from consolidation and the investment carrying value as at September 30, 2007 are detailed as follows:

CASH	$15,378
EXPLORATION ADVANCES AND OTHER RECEIVABLES	171
OTHER ASSETS	75
PROPERTY, PLANT AND EQUIPMENT	146
RESOURCE PROPERTIES	5,774
	21,544
ACCOUNTS PAYABLE AND ACCRUED CHARGES	(662)
NON-CONTROLLING INTEREST	(10,360)
INVESTMENT AS AT APRIL 1, 2007	$10,522
SALE OF SHARES	(136)
EQUITY IN LOSS (APRIL 1, 2007 TO SEPTEMBER 30, 2007)	(755)
GAIN ON DILUTION	74
INVESTMENT AS AT SEPTEMBER 30, 2007	$9,705

- Presence of no data between tables.

The Company presents functional consolidated statements of operations and comprehensive loss in which expenses are aggregated according to the function to which they relate. The Company has identified the major functions as selling, general and administrative expenses; research and development expenses; and patent litigation and reexamination expenses. The following tables present the expenses based on their nature:

2015	Selling, general and administrative	Research and development	Patent litigation and reexamination	Total
Salaries, contractors, commissions and benefits	$ 271,349	$ 668,528	$ –	$ 939,877
Stock-based compensation	293,410	–	–	293,410
Patent litigation-related expenses	–	–	15,254	15,254
Other operating expenses	269,765	115,791	–	385,556
	$ 834,524	$ 784,319	$ 15,254	$ 1,634,097

2014	Selling, general and administrative	Research and development	Patent litigation and reexamination	Total
Salaries, contractors, commissions and benefits	$ 438,392	$ 662,554	$ –	$ 1,100,946
Stock-based compensation	514,663	–	–	514,663
Patent litigation-related expenses	–	–	267,968	267,968
Other operating expenses	374,584	101,184	–	475,768
	$ 1,327,639	$ 763,738	$ 267,968	$ 2,359,345

Challenges

- Column header hierarchies.

	Ownership %	September 30, 2007 Carrying Value	Quoted Market Value
ZINCORE METALS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	49.3	$9,705	$22,860
OTHER INVESTMENTS	--	3,482	3,482
		$13,187	$26,342

	Ownership %	December 31, 2006 Carrying Value	Quoted Market Value
SUPERIOR DIAMONDS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	14.8	$1,919	$3,082
OTHER INVESTMENTS	--	2,485	5,913
		$4,404	$8,995

In April 2007 the Company sold 500,000 common shares of Zincore Metals Inc. ("Zincore") for gross proceeds of $350,000 and recorded a gain of $212,000. This sale caused the Company's interest in Zincore to be reduced from 50.4% to 49.7%. As a result of the reduction of the Company's interest and resulting loss of control, the assets and liabilities of Zincore were no longer consolidated in the Company's balance sheet effective April 1, 2007. The assets and liabilities of Zincore excluded from consolidation and the investment carrying value as at September 30, 2007 are detailed as follows:

CASH	$15,378
EXPLORATION ADVANCES AND OTHER RECEIVABLES	171
OTHER ASSETS	75
PROPERTY, PLANT AND EQUIPMENT	146
RESOURCE PROPERTIES	5,774
	21,544
ACCOUNTS PAYABLE AND ACCRUED CHARGES	(662)
NON-CONTROLLING INTEREST	(10,360)
INVESTMENT AS AT APRIL 1, 2007	$10,522
SALE OF SHARES	(136)
EQUITY IN LOSS (APRIL 1, 2007 TO SEPTEMBER 30, 2007)	(755)
GAIN ON DILUTION	74
INVESTMENT AS AT SEPTEMBER 30, 2007	$9,705

- Row header hierarchies.

01 COMMUNIQUE LABORATORY INC.

Consolidated Statements of Cash Flows
(In Canadian dollars)

Years ended October 31, 2015 and 2014

	2015	2014
Cash provided by (used in):		
Operating activities:		
Loss for the year	$ (1,571,724)	$ (2,073,332)
Adjustments to reconcile loss for the year to net cash flows from operating activities:		
Depreciation of property and equipment	6,112	7,488
Stock-based compensation expense	293,410	514,663
Accretion on liability portion of debenture	6,825	–
Interest paid on debenture	20,650	–
Interest income	(2,863)	(19,322)
Change in non-cash operating working capital (note 8)	(157,073)	(223,666)
	(1,404,663)	(1,794,169)
Interest income received	2,863	19,322
	(1,401,800)	(1,774,847)
Financing activities:		
Proceeds from the exercise of stock options and agent options	205,000	75,600
Proceeds from issuance of a debenture (note 5)	400,000	–
Interest paid on debenture	(20,650)	–
	584,350	75,600

Challenges

- Missing contextual information

	Ownership %	September 30, 2007 Carrying Value	Quoted Market Value
ZINCORE METALS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	49.3	$9,705	$22,860
OTHER INVESTMENTS	--	3,482	3,482
		$13,187	$26,342

	Ownership %	December 31, 2006 Carrying Value	Quoted Market Value
SUPERIOR DIAMONDS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	14.8	$1,919	$3,082
OTHER INVESTMENTS	--	2,485	5,913
		$4,404	$8,995

In April 2007 the Company sold 500,000 common shares of Zincore Metals Inc. ("Zincore") for gross proceeds of $350,000 and recorded a gain of $212,000. This sale caused the Company's interest in Zincore to be reduced from 50.4% to 49.7%. As a result of the reduction of the Company's interest and resulting loss of control, the assets and liabilities of Zincore were no longer consolidated in the Company's balance sheet effective April 1, 2007. The assets and liabilities of Zincore excluded from consolidation and the investment carrying value as at September 30, 2007 are detailed as follows:

CASH	$ 5,378
EXPLORATION ADVANCES AND OTHER RECEIVABLES	171
OTHER ASSETS	75
PROPERTY, PLANT AND EQUIPMENT	146
RESOURCE PROPERTIES	5,774
	21,544
ACCOUNTS PAYABLE AND ACCRUED CHARGES	(662)
NON-CONTROLLING INTEREST	(10,360)
INVESTMENT AS AT APRIL 1, 2007	$ 0,522
SALE OF SHARES	(136)
EQUITY IN LOSS (APRIL 1, 2007 TO SEPTEMBER 30, 2007)	(755)
GAIN ON DILUTION	74
INVESTMENT AS AT SEPTEMBER 30, 2007	$9,705

- Missing Aggregate type row headers

	Ownership %	September 30, 2007 Carrying Value	Quoted Market Value
ZINCORE METALS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	49.3	$9,705	$22,860
OTHER INVESTMENTS	--	3,482	3,482
		$13,187	$26,342

	Ownership %	December 31, 2006 Carrying Value	Quoted Market Value
SUPERIOR DIAMONDS INC. – SIGNIFICANTLY INFLUENCED AFFILIATE	14.8	$1,919	$3,082
OTHER INVESTMENTS	--	2,485	5,913
		$4,404	$8,995

In April 2007 the Company sold 500,000 common shares of Zincore Metals Inc. ("Zincore") for gross proceeds of $350,000 and recorded a gain of $212,000. This sale caused the Company's interest in Zincore to be reduced from 50.4% to 49.7%. As a result of the reduction of the Company's interest and resulting loss of control, the assets and liabilities of Zincore were no longer consolidated in the Company's balance sheet effective April 1, 2007. The assets and liabilities of Zincore excluded from consolidation and the investment carrying value as at September 30, 2007 are detailed as follows:

CASH	$15,378
EXPLORATION ADVANCES AND OTHER RECEIVABLES	171
OTHER ASSETS	75
PROPERTY, PLANT AND EQUIPMENT	146
RESOURCE PROPERTIES	5,774
	21,544
ACCOUNTS PAYABLE AND ACCRUED CHARGES	(662)
NON-CONTROLLING INTEREST	(10,360)
INVESTMENT AS AT APRIL 1, 2007	$10,522
SALE OF SHARES	(136)
EQUITY IN LOSS (APRIL 1, 2007 TO SEPTEMBER 30, 2007)	(755)
GAIN ON DILUTION	74
INVESTMENT AS AT SEPTEMBER 30, 2007	$9,705

Proposed Approach

Treat tables and cells as objects.

We can leverage Objection Detection Deep Learning works to detect tables and cells.

Challenge:

How do we get the data?

1. There are some amount of data which has annotation for table boundaries, but not for cell boundaries.
2. Scientific papers are often present in HTML and PDF. HTML codes are structured. The authors did token matching to annotate 568k scientific tables from Pubmed dataset.

Proposed Approach (continued)

Can we simply finetune existing object detectors with the newly created data?

-> Tables and cells have very different aspect ratios.

What’s the intuition:

Cells are always inside tables and tables always contain cells.

Proposed Approach (continued)

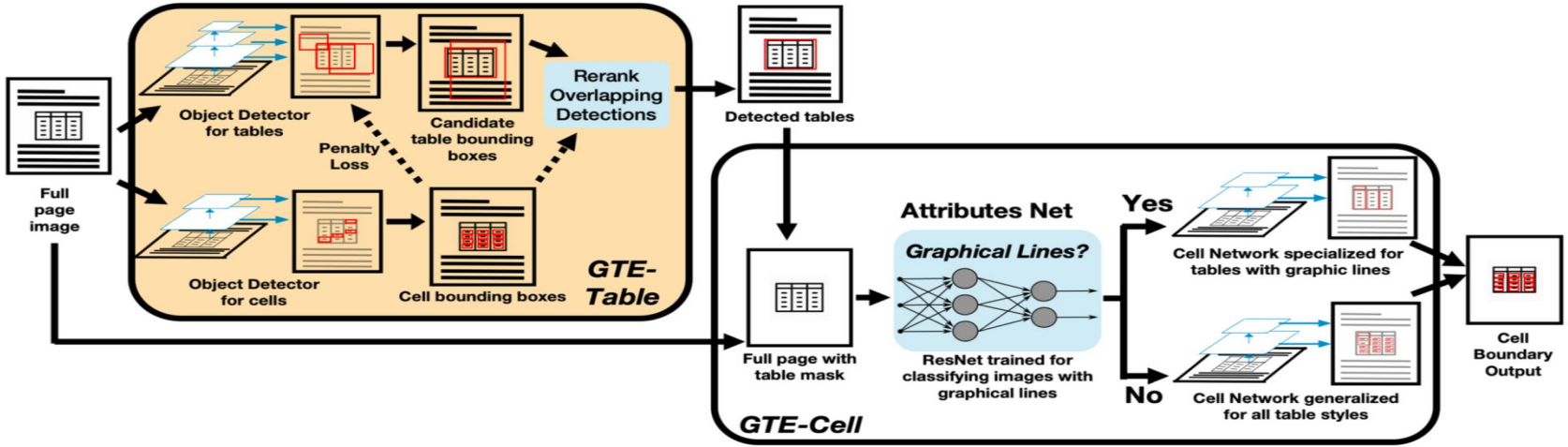


Figure 2: Our full GTE Framework consists of several networks for table (GTE-Table) and cell (GTE-Cell) boundary detection. The input is an image form of a document page for both sub-frameworks, but note GTE-Cell depends on table boundaries output by GTE-Table to generate cell structures for each specific table.

Table Detection Framework

An object detector for tables and an object detector for cells connected with a penalty loss.

The cell detector detects how dense the cells are. The predicted tables which do not contain a lot of cells gets high penalty.

During inference time, usual object detection methods use max-suppression for overlapping candidates (take the one with the highest confidence, suppress others)

This work introduces tableness criteria, based on the density of cells inside and outside of the table to rerank tables that overlap with similar confidences.

Outputs table boundaries.

Table Detection Framework: Penalty Loss

A penalty on table classification is applied if:

1. The detection is a table but contains very little cell mask inside the bounding box

2. The detection is not a table but contains a lot of cells inside the bounding box

Table Detection Framework: Reranking Candidates

Re-rank:

When tables are overlapping and have similar confidence, select the table based on Table-ness:

1. Should have lots of cellular regions
2. Should not have cellular regions just outside that is not being covered by a non-overlapping region

Performance for Table Detection

State of the art on ICDAR 2013 and ICDAR 2019 datasets

ICDAR 2013

Character level recall, precision and F1

Category	Method	Input type	Recall	Precision	F1
Commercial Softwares	*FineReader*	PDF	99.71	97.29	98.48
Non Deep Learning	*Nurminen*[8]	PDF	90.77	92.10	91.43
Deep Learning	*TableBank*[18]	Image	/	/	96.25
Ours	GTE	Image	**99.77**	**98.97**	**99.31**
Ablation	Detection-Base	Image	84.64	90.65	84.65
	GTE-Table-Sep	Image	95.71	98.18	95.71

ICDAR 2019

Precision and Recall at different levels of Intersection Over Union (IOU)

Method	IOU = 0.8		IOU = 0.9		Weighted F1
	P	R	P	R	
NLPR-PAL[4]	93	93	86	86	93
TableRadar[4]	95	94	**90**	**89**	**94**
GTE	**96**	**95**	**90**	**89**	**94**

Cell Detection framework

Object detection models typically focus on local areas, don’t focus on global styles

Cell Detection framework:

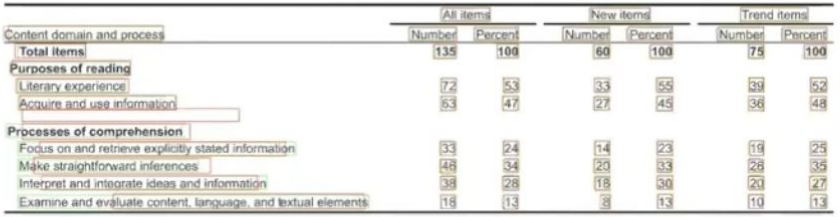
Attributes net, classifies image whether a table has graphical lines or not. (one of the styles this paper focused on), two cell networks specialized for whether graphical lines are present or not.

Even if we have cell locations, we still don’t know which cells are in same row or in the same column or not.

Solution:

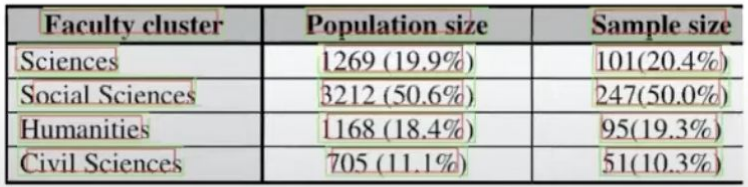
Use the centers of the predicted cell boxes as cluster centers and utilize a k-means algorithm to determine the row and column number for every cell.

Detection to structure with location and alignment clustering



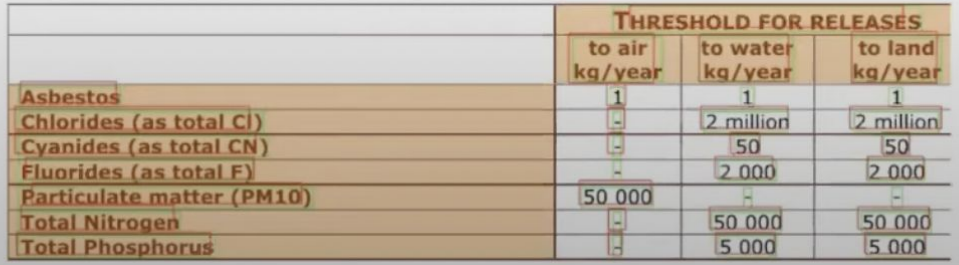
Content domain and process	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
Total items	135	100	60	100	75	100
Purposes of reading						
Literary experience	72	53	33	55	39	52
Acquire and use information	63	47	27	45	36	48
Processes of comprehension						
Focus on and retrieve explicitly stated information	33	24	14	23	19	25
Make straightforward inferences	48	34	20	33	28	35
Interpret and integrate ideas and information	38	28	18	30	20	27
Examine and evaluate content, language, and textual elements	18	13	8	13	10	13

Content domain and process	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
Total items	135	100	60	100	75	100
Purposes of reading						
Literary experience	72	53	33	55	39	52
Acquire and use information	63	47	27	45	36	48
Processes of comprehension						
Focus on and retrieve explicitly stated information	33	24	14	23	19	25
Make straightforward inferences	46	34	20	33	26	35
Interpret and integrate ideas and information	38	28	18	30	20	27
Examine and evaluate content, language, and textual elements	18	13	8	13	10	13



Faculty cluster	Population size	Sample size
Sciences	1269 (19.9%)	101(20.4%)
Social Sciences	3212 (50.6%)	247(50.0%)
Humanities	1168 (18.4%)	95(19.3%)
Civil Sciences	705 (11.1%)	51(10.3%)

Faculty cluster	Population size	Sample size
Sciences	1269 (19.9%)	101(20.4%)
Social Sciences	3212 (50.6%)	247(50.0%)
Humanities	1168 (18.4%)	95(19.3%)
Civil Sciences	705 (11.1%)	51(10.3%)



	THRESHOLD FOR RELEASES		
	to air kg/year	to water kg/year	to land kg/year
Asbestos	1	1	1
Chlorides (as total Cl)	-	2 million	2 million
Cyanides (as total CN)	-	50	50
Fluorides (as total F)	-	2 000	2 000
Particulate matter (PM10)	50 000	-	-
Total Nitrogen	-	50 000	50 000
Total Phosphorus	-	5 000	5 000

	THRESHOLD FOR RELEASES		
	to air kg/year	to water kg/year	to land kg/year
Asbestos	1	1	1
Chlorides (as total Cl)	-	2 million	2 million
Cyanides (as total CN)	-	50	50
Fluorides (as total F)	-	2 000	2 000
Particulate matter (PM10)	50 000	-	-
Total Nitrogen	-	50 000	50 000
Total Phosphorus	-	5 000	5 000

Cell Structure Metric

ICDAR 2013, 2019 Competition

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(a) Original table as in ground truth

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(b) Incorrectly recognized cell structure with split column

■ Correct adjacency relations □ Incorrect adjacency relations

$$\text{Recall} = \frac{\text{correct adjacency relations}}{\text{total adjacency relations}} = \frac{24}{31} = 77.4\%$$

$$\text{Precision} = \frac{\text{correct adjacency relations}}{\text{detected adjacency relations}} = \frac{24}{28} = 85.7\%$$

Cell Structure Metric

Recall and Precision of Cell Adjacency Relations

-ICDAR2013 -> Match based on text

-ICDAR2019 -> Match based on bounding box IOU

Performance for Cell Structure Recognition in ICDAR Competitions

State of the art for Cell Structure Recognition in ICDAR 2013 +2019 Datasets

ICDAR 2013

Method	GT?	Rec.	Prec.	F1
Nurminen[8]	N	80.78	86.93	83.74
GTE	N	**92.72**	**94.41**	**93.50**
Tensmeyer[33]	Y	94.64	95.89	95.26
GTE	Y	**95.77**	**96.76**	**96.24**
Detection-Base	Y	76.66	80.63	78.10
GTE-Cell-Style-Mix -no-pt	Y	89.78	89.30	89.43
GTE-Cell-Style-Mix	Y	92.39	94.20	93.15
GTE-Cell-Border	Y	91.60	93.67	92.48

ICDAR 2019

Method	IOU			Weighted F1
	0.1	0.5	0.6	
NLPR-PAL[4]	-	36.5	30.5	20.6
CascadeTabNet[23]	-	43.8	35.4	23.2
GTE	77.5	**54.8**	**38.5**	**24.8**

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

Task Description

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.

The authors address 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).

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.

Existing Works and Datasets

Popular question answering (e.g. SQuAD and Natural Question (Rajpurkar et al., 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 closest 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, (close to the “unknown” statements in this work.)

Why a New Dataset?

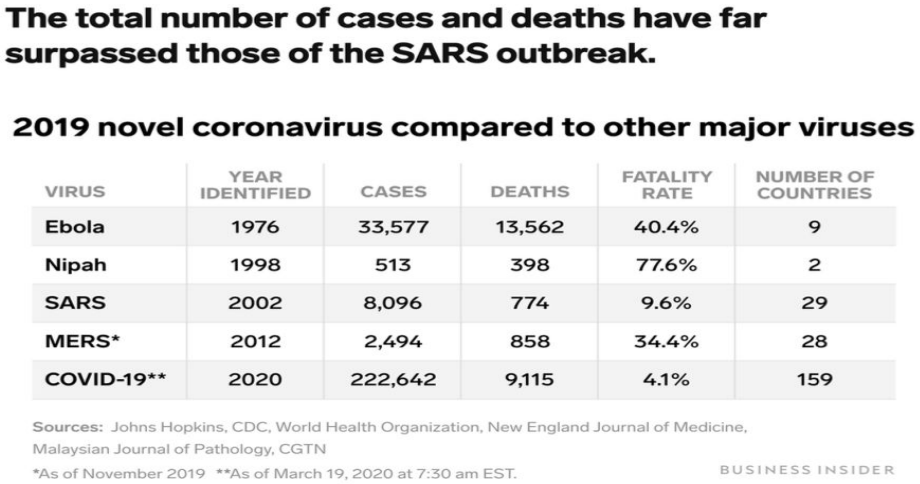
Scientific tables have very specialized vocabulary and can be more difficult to interpret. Additionally, scientific tables have much more complex structure, like hierarchical column and row headers, rendering the assumption that the first column/row is the header unhelpful. Finally, tables are often directly referenced in scientific text unlike Wikipedia tables that are generally stand-alone.

The second key differentiator of SEM-TAB-FACTS is the accompanying evidence annotations.

Sample Tables

	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.



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.

Sample Example

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

Original Generated Statements

Entailed

- *There are 9 different types country in the given table.*
- *The n value is same for Hong Kong and South Korea.*
- *There are 4 different types of Networks which contains its own mean, SD and max.*
- *The least max value is 0 in Political network of Taiwan.*
- *All the values of SD in Board network is greater than the values of SD in Family network.*

Refuted

- *All the values of SD in Board network is less than the values of SD in Family network.*
- *There are 4 different types of Networks which contains same mean, SD and max.*
- *The least max value is 0 in Political network of Thailand.*
- *There are 7 different types country in the given table.*
- *The n value is same for Hong Kong and Malaysia.*

Original Related Natural In-text Statements

- *Descriptive statistics for each board network type are offered in Table 2 , broken down by country.*
- *For each network interaction, there is considerable variation both across and within countries.*

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.

Dataset Splits

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.

Evaluation Metrics (Task A: Statement Fact Verification)

2 evaluation metrics:

1. 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.

2. A 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/ en- tailed statement as unknown.

Evaluation Metrics (Task B: Evidence Detection)

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.

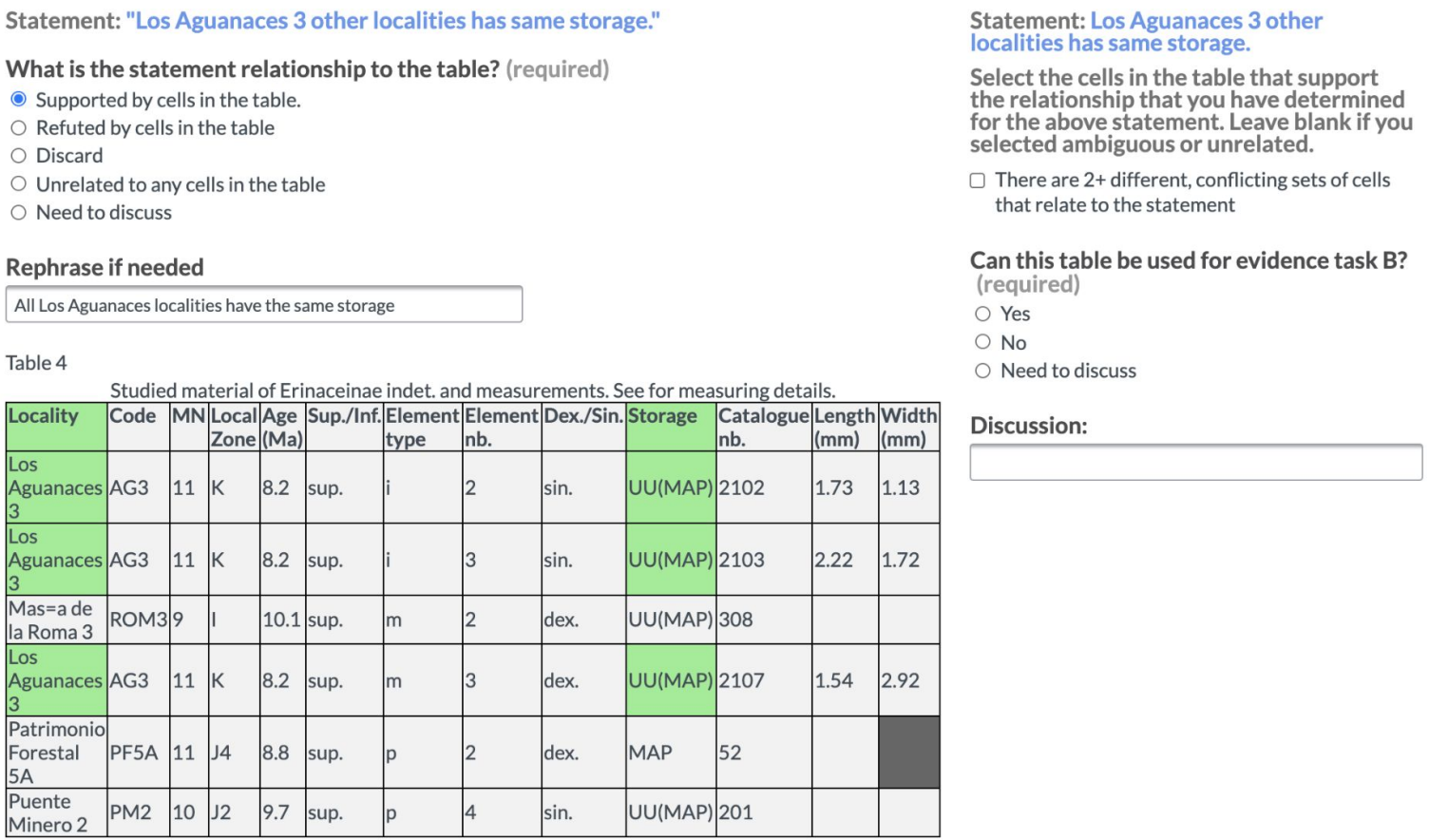


Table 4

Studied material of Erinaceinae indet. and measurements. See for measuring details.

Locality	Code	MN	Local Zone	Age (Ma)	Sup./Inf.	Element type	Element nb.	Dex./Sin.	Storage	Catalogue nb.	Length (mm)	Width (mm)
Los Aguanaces 3	AG3	11	K	8.2	sup.	i	2	sin.	UU(MAP)	2102	1.73	1.13
Los Aguanaces 3	AG3	11	K	8.2	sup.	i	3	sin.	UU(MAP)	2103	2.22	1.72
Mas=a de la Roma 3	ROM3	9	I	10.1	sup.	m	2	dex.	UU(MAP)	308		
Los Aguanaces 3	AG3	11	K	8.2	sup.	m	3	dex.	UU(MAP)	2107	1.54	2.92
Patrimonio Forestal 5A	PF5A	11	J4	8.8	sup.	p	2	dex.	MAP	52		
Puente Minero 2	PM2	10	J2	9.7	sup.	p	4	sin.	UU(MAP)	201		

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

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.

Leaderboard

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

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

Team Approaches for Task A

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.

Team Approaches for Task B

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)

Thank you