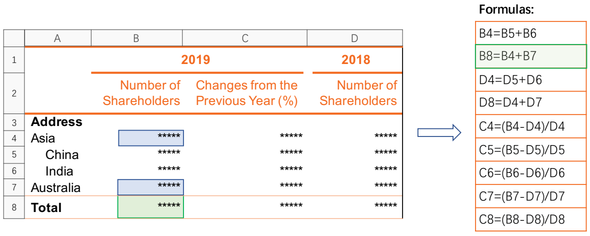
Numerical Formula Recognition from Tables

Qingping Yang[1,2], Yixuan Cao[1,2], Hongwei Li[3], Ping Luos[1,2,4]

[1] Institute of Computing Technology, Chinese Academy of Sciences, [2] University of Chinese Academy of Sciences
[3] Research Department, P.A.I. Ltd., [4] Peng Cheng Laboratory

Introduction

- Claims over the numerical relationships among some measures are commonly expressed as formulas in tabular forms
- This paper introduces the problem of numerical formula recognition from tables



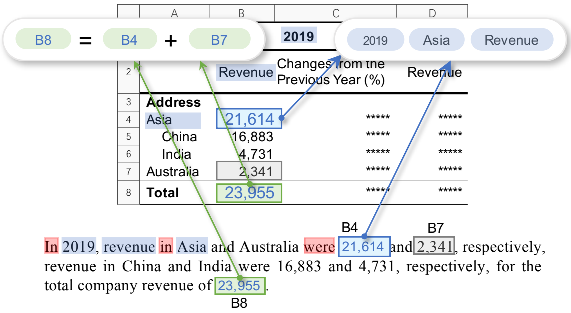
	A	B	C	D
1		2019		2018
2		Number of Shareholders	Changes from the Previous Year (%)	Number of Shareholders
3	**Address**			
4	Asia	*****	*****	*****
5	China	*****	*****	*****
6	India	*****	*****	*****
7	Australia	*****	*****	*****
8	**Total**	*****	*****	*****

Formulas:

B4=B5+B6
B8=B4+B7
D4=D5+D6
D8=D4+D7
C4=(B4-D4)/D4
C5=(B5-D5)/D5
C6=(B6-D6)/D6
C7=(B7-D7)/D7
C8=(B8-D8)/D8

Rethinking on Table

- Table is a kind of **language** that adopts a different linguistic paradigm from natural language.
- *Content words* are scattered regularly in table cells, and *visual grammar* express the grammatical relationships among the table cells.

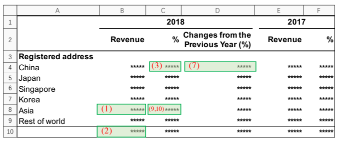


	A	B	C	D
1		2019		
2		Revenue	Changes from the Previous Year (%)	Revenue
3	**Address**			
4	Asia	21,614	*****	*****
5	China	16,883	*****	*****
6	India	4,731	*****	*****
7	Australia	2,341	*****	*****
8	**Total**	23,955	*****	*****

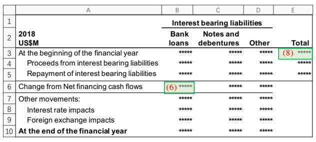
In 2019, revenue in Asia and Australia were 21,614 and 2,341, respectively, revenue in China and India were 16,883 and 4,731, respectively, for the total company revenue of 23,955.

Challenges

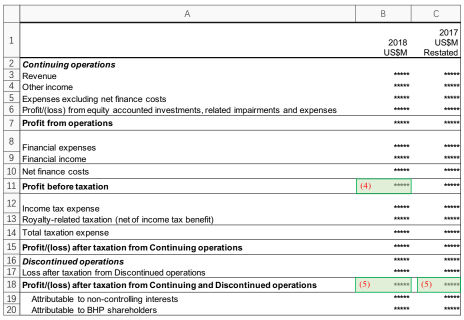
- Recognizing formulas require decoding the visual grammar while simultaneously understanding the textual information.
- Horizontal formulas are common in tables.
- Multiple formulas might appear in the same table cell.
- Formula Complexity



	A	B	C	D	E	F
1		2018			2017	
2		Revenue	%	Changes from the Previous Year (%)	Revenue	%
3	**Registered address**					
4	China	*****	(3) *****	(7) *****	*****	*****
5	Japan	*****	*****	*****	*****	*****
6	Singapore	*****	*****	*****	*****	*****
7	Korea	*****	*****	*****	*****	*****
8	Asia	(1) *****	(9,10) *****	*****	*****	*****
9	Rest of world	*****	*****	*****	*****	*****
10		(2) *****	*****	*****	*****	*****



	A	B	C	D	E
1		Interest bearing liabilities			
2	**2018** **US$M**	Bank loans	Notes and debentures	Other	Total
3	At the beginning of the financial year	*****	*****	*****	(8) *****
4	Proceeds from interest bearing liabilities	*****	*****	*****	*****
5	Repayment of interest bearing liabilities	*****	*****	*****	*****
6	Change from Net financing cash flows	(6) *****	*****	*****	
7	Other movements:	*****	*****	*****	
8	Interest rate impacts	*****	*****	*****	
9	Foreign exchange impacts	*****	*****	*****	
10	**At the end of the financial year**	*****	*****	*****	



	A	B	C
1		2018 US$M	2017 US$M Restated
2	***Continuing operations***		
3	Revenue	*****	*****
4	Other income	*****	*****
5	Expenses excluding net finance costs	*****	*****
6	Profit/(loss) from equity accounted investments, related impairments and expenses	*****	*****
7	**Profit from operations**	*****	*****
8	Financial expenses	*****	*****
9	Financial income	*****	*****
10	Net finance costs	*****	*****
11	**Profit before taxation**	(4) *****	*****
12	Income tax expense	*****	*****
13	Royalty-related taxation (net of income tax benefit)	*****	*****
14	Total taxation expense	*****	*****
15	**Profit/(loss) after taxation from Continuing operations**	*****	*****
16	***Discontinued operations***	*****	*****
17	Loss after taxation from Discontinued operations	*****	*****
18	**Profit/(loss) after taxation from Continuing and Discontinued operations**	(5) *****	(5) *****
19	Attributable to non-controlling interests	*****	*****
20	Attributable to BHP shareholders	*****	*****

Methods - TaFor

- Problem Conversion.

A formula can be defined as:

$$r = f(e_1, \dots, e_i, \dots, e_n)$$

Converted to a set of triplets as:

$$\{(r, f^1, e_1), \dots, (r, f^i, e_i), \dots, (r, f^i, e_i)\},$$

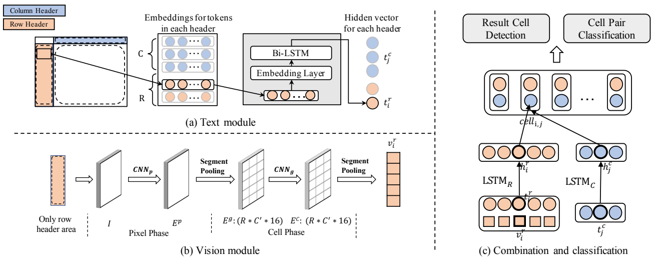
where r is the result cell, f is the formula type, e is the element cell.

- Two Steps.

1. Result Cell Detection
2. Cell Pair Classification

How to encode a table and cell inside it?

- Two-channel Model



Experiments

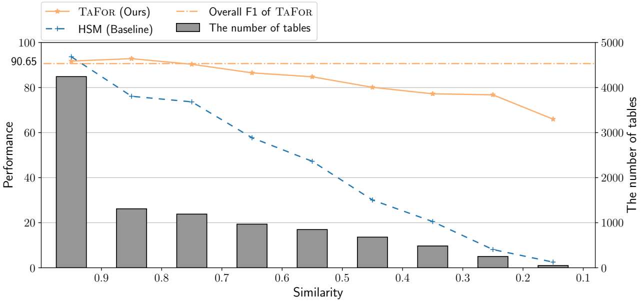
Table 2: Evaluation results.

	±	*d*	*gr*	*avg*	overall
HHM	42.57	46.29	48.78	46.37	44.08
HSM	68.00	78.97	74.45	67.12	72.05
TaFor	**90.15**	91.66	85.87	87.38	90.65
HHM + TaFor	90.02	**93.58**	**92.19**	**89.18**	**91.31**

Table 4: Ablation results.

	Result cell detection	Pair level	Formula level				
			±	*d*	*gr*	*avg*	overall
TaFor	96.12	95.17	90.15	91.66	85.87	87.38	90.65
−text	61.43	65.42	64.24	0	0	46.40	48.78
−vision	94.42	93.93	87.86	90.89	83.69	83.59	88.77

The generalization ability of TaFor


TaFor (Ours)
HSM (Baseline)
Overall F1 of TaFor
The number of tables
Performance
Similarity
The number of tables