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.

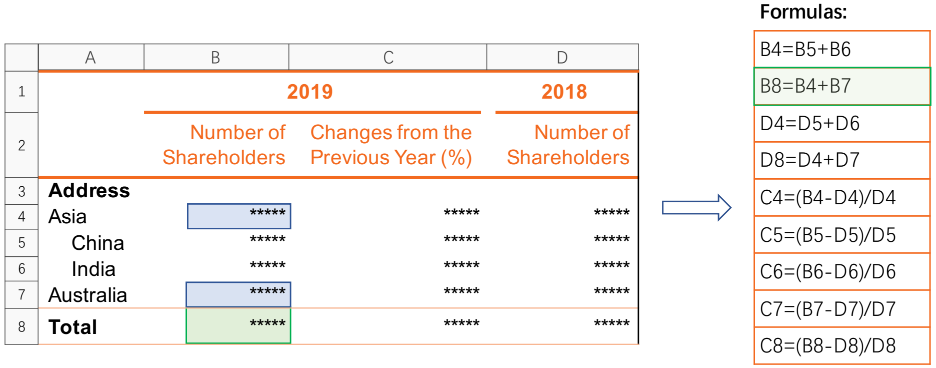
[4] Peng Cheng Laboratory

Background

- Claims over the numerical relationships among some objective measures widely exist in the published documents on the Web.

- These numerical relationships are often expressed in tabular forms.

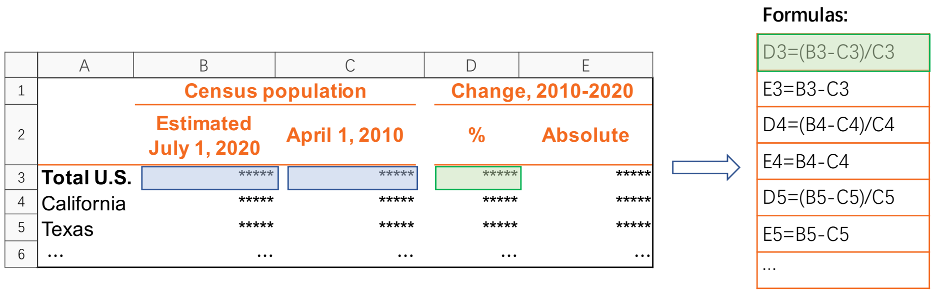
- Task: **Numerical Formula Recognition (NFR) from tables**



	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



	A	B	C	D	E
1		Census population		Change, 2010-2020	
2		Estimated July 1, 2020	April 1, 2010	%	Absolute
3	Total U.S.	*****	*****	*****	*****
4	California	*****	*****	*****	*****
5	Texas	*****	*****	*****	*****
6	…	…	…	…	…

Formulas:

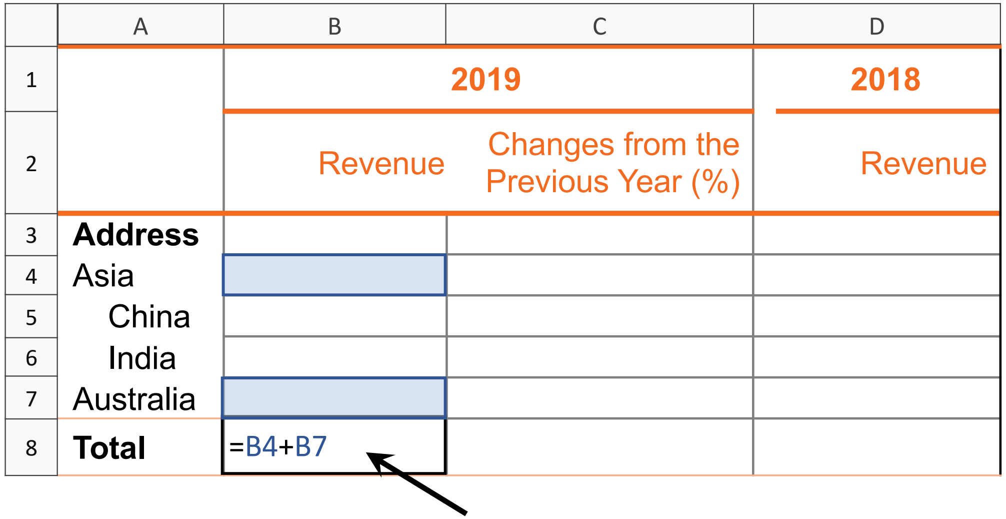
D3=(B3-C3)/C3
E3=B3-C3
D4=(B4-C4)/C4
E4=B4-C4
D5=(B5-C5)/C5
E5=B5-C5
…

Application

- *Error Correction in Tables*
  - Numerical errors caused by formulas are inevitable, even in published documents which have been reviewed many times.
  - These errors may cause severe consequences.
    - 2012, JP Morgan suffered $6.5 billion in losses and fines.
    - 2013, the paper “Growth in a Time of Debt” led to unjustified austerity policies.

Application

- *Formula Recommendation in Tables*
  - After users have filled in the table headers and overall table layout is developed, we can automatically suggest the formulas among table cell.



	A	B	C	D
1		**2019**		**2018**
2		Revenue	Changes from the Previous Year (%)	Revenue
3	**Address**			
4	Asia			
5	China			
6	India			
7	Australia			
8	**Total**	=B4+B7		

Formula Recommendation

Challenges

- Numerical values and existing formulas are not reliable.
  - Values in tables are error-prone. [1, 2]
  - At least one error caused by a formula was found in more than 95% of spreadsheets. [3]

- Need a more reliable method.

[1] WARDER: Refining cell clustering for effective spreadsheet defect detection via validity properties. 2019.
[2] A critical review of the literature on spreadsheet errors. 2008.
[3] What we don't know about spreadsheet errors today: The facts, why we don't believe them, and what we need to do. 2016.

Challenges

- Formula complexity
  - A formula in table can be define as:

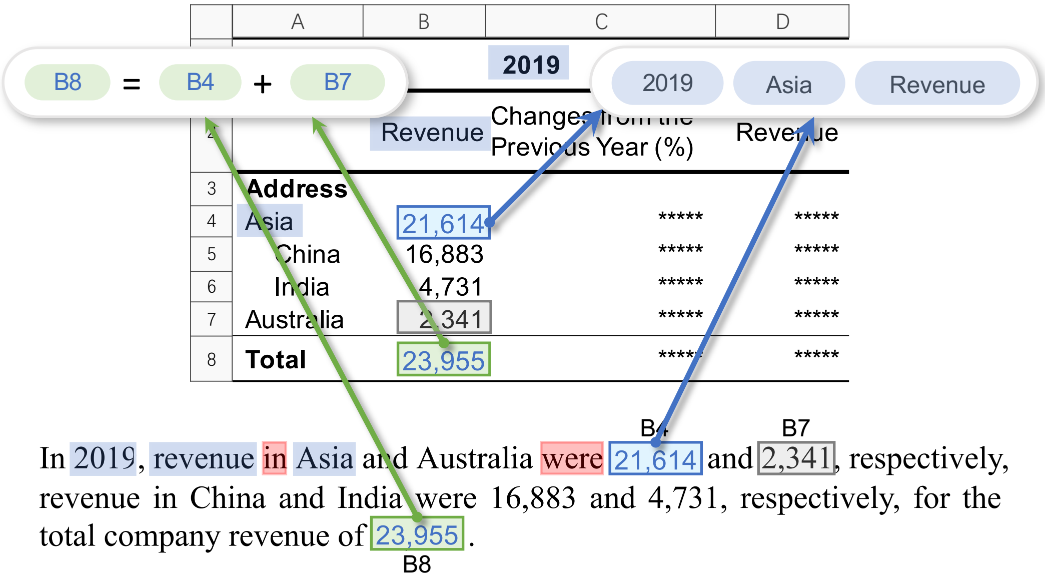
$$r = f(e_1, \cdots, e_i, \cdots, e_n)$$

  - For example $r = e_1/e_2$ can be expressed as $r = f_{div}(e_1, e_2)$.

1. Diverse math function.
2. The number of arguments cannot be fixed in advance (e.g. SUM).
3. The order of arguments (e.g. division).
4. Commutative property (e.g. SUM, AVG, MIN, MAX)

Challenges

- Table representation complexity
  - Table is a kind of ***language*** that adopts a different linguistic paradigm from natural language.

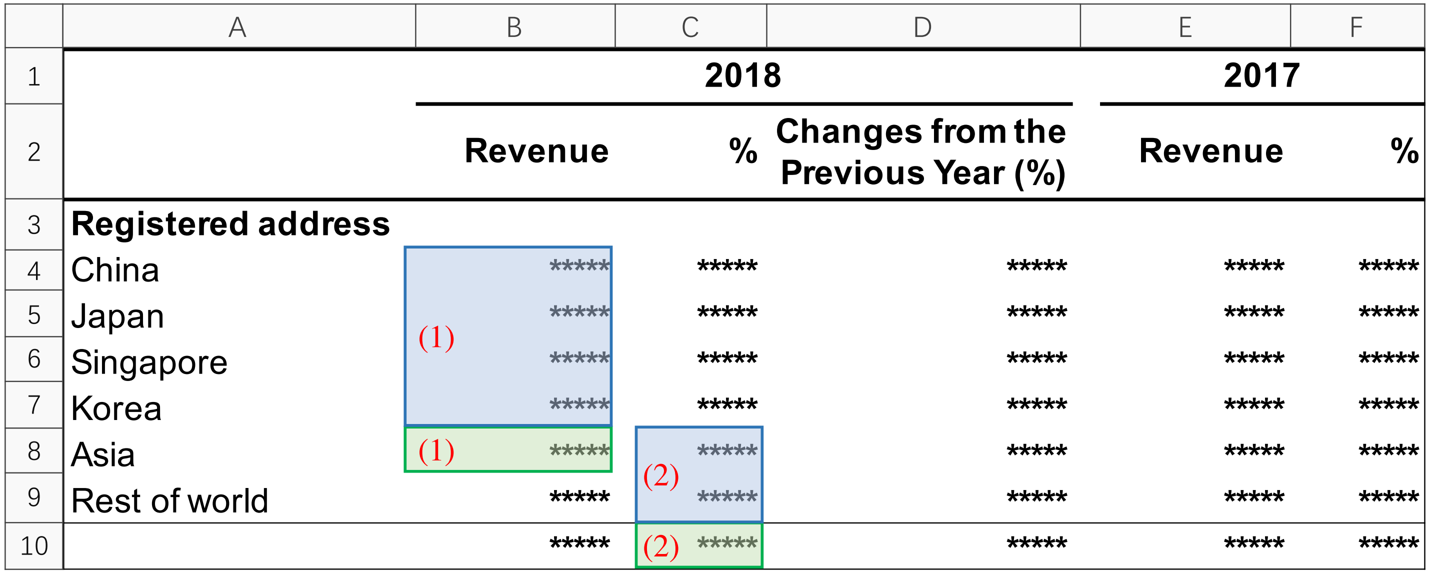


	A	B	C	D
			2019	
2		Revenue	Changes from the Previous Year (%)	Revenue
3	Address			
4	Asia	21,614	*****	*****
5	China	16,883	*****	*****
6	India	4,731	*****	*****
7	Australia	2,341	*****	*****
8	Total	23,955	*****	*****

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

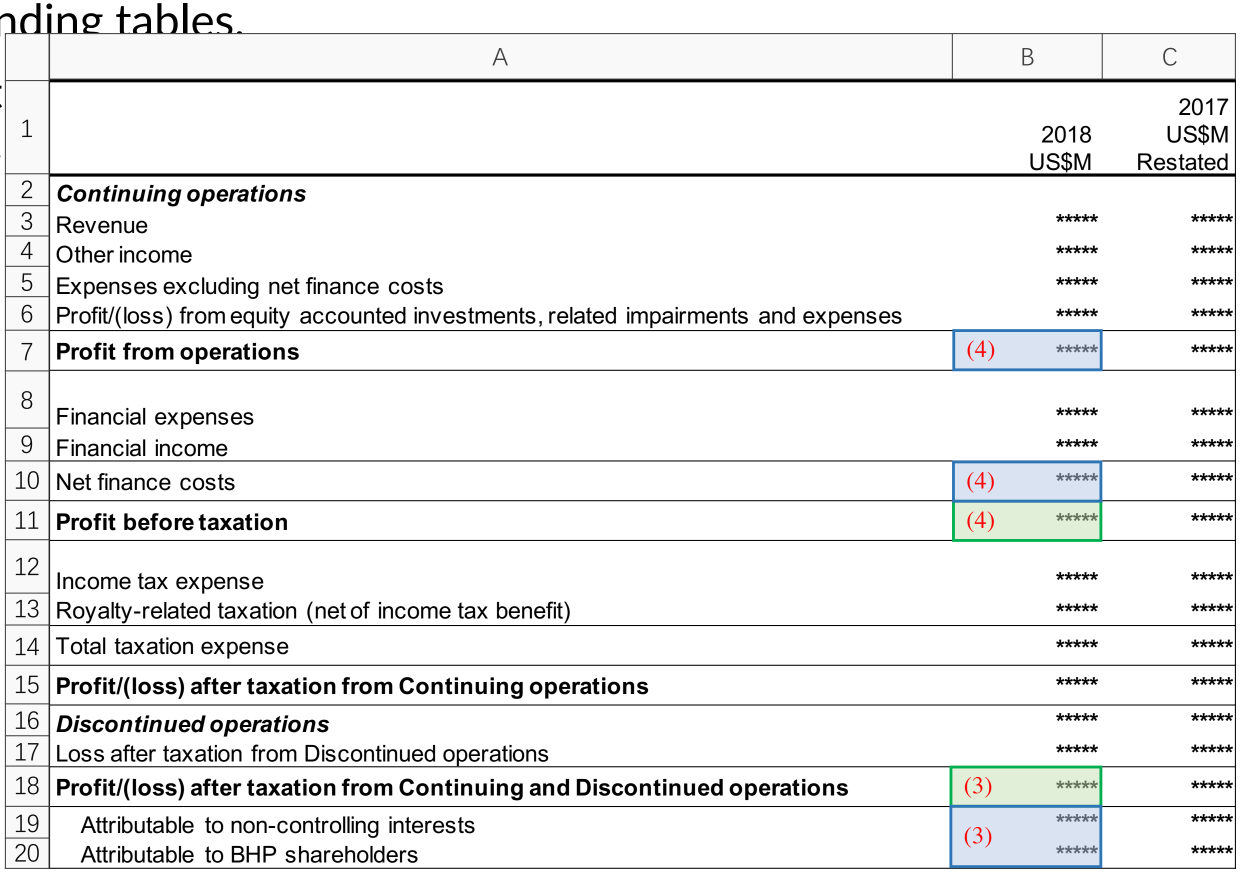
- Table representation complexity
  - Observation 1: Textual information on the header hierarchy is the key to understanding tables.



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

Challenges

- Table representation complexity
  - Observation 1: Textual information on the header hierarchy is the key to understanding tables.
  - Observat r representing formulas.



	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**	(4) *****	*****
8	Financial expenses	*****	*****
9	Financial income	*****	*****
10	Net finance costs	(4) *****	*****
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**	(3) *****	*****
19	Attributable to non-controlling interests	(3) *****	*****
20	Attributable to BHP shareholders	*****	*****

Challenges

- Table representation complexity
  - Observation 1: Textual information on the header hierarchy is the key to understanding tables.
  - Observation 2: The visual appearances serve as auxiliary information for representing formulas.
  - Observation 3: Horizontal formulas are common in tables.
  - Observation 4: Multiple Formulas might appear in the same table cell.

Solution Overview

- The formula recognition task → a relation extraction task between two cells
    - by first detect result cells and then classify cell pairs.

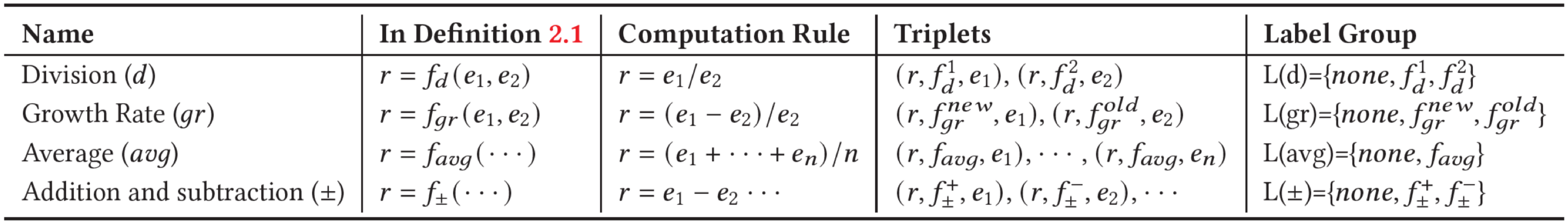
- To do the classification, a table cell encoding model TAFOR is proposed which considers both textual and visual information.

- We leverage the text and visual appearance of table headers and table layout structure, which are more reliable features.

Solution – Problem Conversion

- Main idea: a formula → several relations between r and e.
- Triplet: (r, f^i, e)
- A formula $r = f(e_1, \cdots, e_i, \cdots, e_n) \to \{(r, f^1, e_1), \cdots, (r, f^i, e_i), \cdots, (r, f^n, e_n)\}$
  - For example, $r = f_{div}(e_1, e_2) \to \{(r, f_{div}^1, e_1), (r, f_{div}^2, e_2)\}$

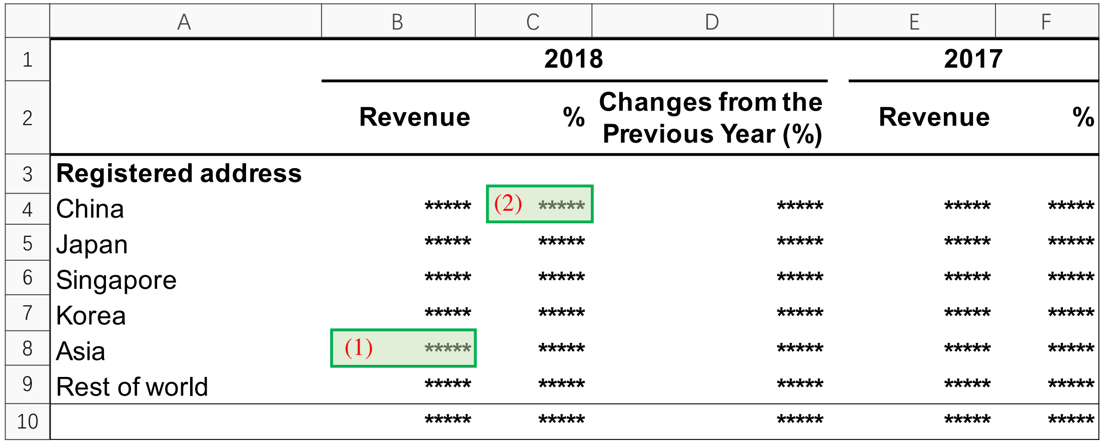
Table 1: Examples of formulas with their triplets.



Name	In Definition 2.1	Computation Rule	Triplets	Label Group
Division (d)	$r = f_d(e_1, e_2)$	$r = e_1/e_2$	$(r, f_d^1, e_1), (r, f_d^2, e_2)$	L(d)=$\{none, f_d^1, f_d^2\}$
Growth Rate (gr)	$r = f_{gr}(e_1, e_2)$	$r = (e_1 - e_2)/e_2$	$(r, f_{gr}^{new}, e_1), (r, f_{gr}^{old}, e_2)$	L(gr)=$\{none, f_{gr}^{new}, f_{gr}^{old}\}$
Average (avg)	$r = f_{avg}(\cdots)$	$r = (e_1 + \cdots + e_n)/n$	$(r, f_{avg}, e_1), \cdots, (r, f_{avg}, e_n)$	L(avg)=$\{none, f_{avg}\}$
Addition and subtraction (±)	$r = f_{\pm}(\cdots)$	$r = e_1 - e_2 \cdots$	$(r, f_{\pm}^{+}, e_1), (r, f_{\pm}^{-}, e_2), \cdots$	L(±)=$\{none, f_{\pm}^{+}, f_{\pm}^{-}\}$

Solution – Framework

1. Result Cell Detection
2. Cell Pair Classification



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

Predicted:

Result cell: B8, C4

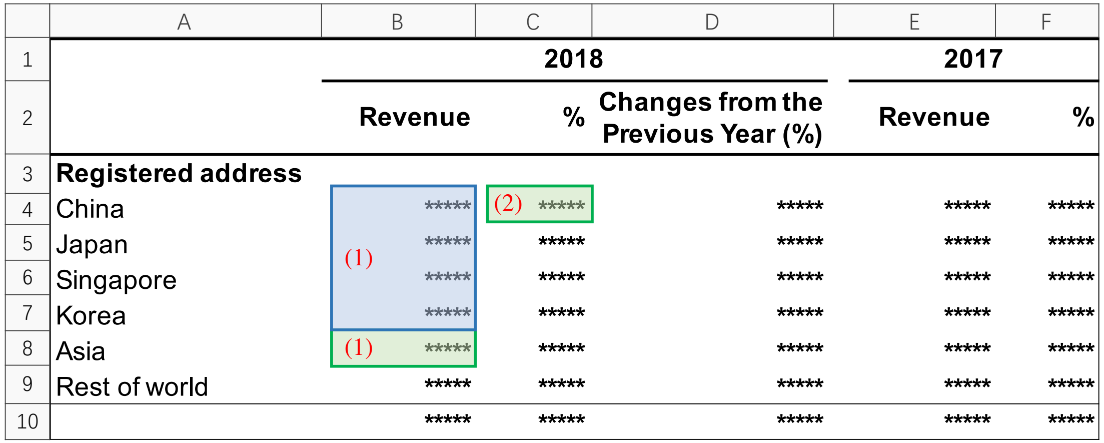
Formula:

B8 =

C4 =

Solution – Framework

1. Result Cell Detection
2. Cell Pair Classification



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

Predicted:

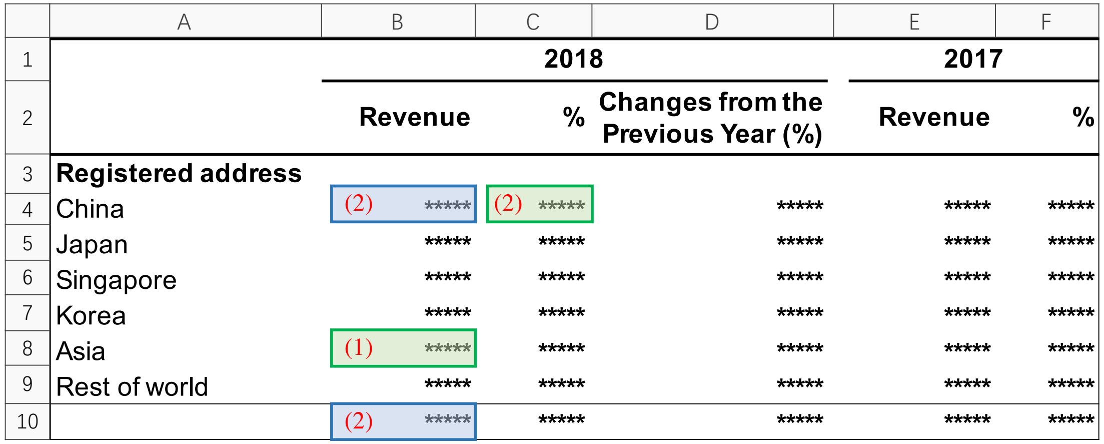
$\{\mathrm{B8}, f_{\pm}, \mathrm{B4}\}, \{\mathrm{B8}, f_{\pm}, \mathrm{B5}\}$
$\{\mathrm{B8}, f_{\pm}, \mathrm{B6}\}, \{\mathrm{B8}, f_{\pm}, \mathrm{B7}\}$

Formula:

B8 = +B4+B5+B6+B7
C4 =

Solution – Framework

1. Result Cell Detection
2. Cell Pair Classification



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

Predicted:

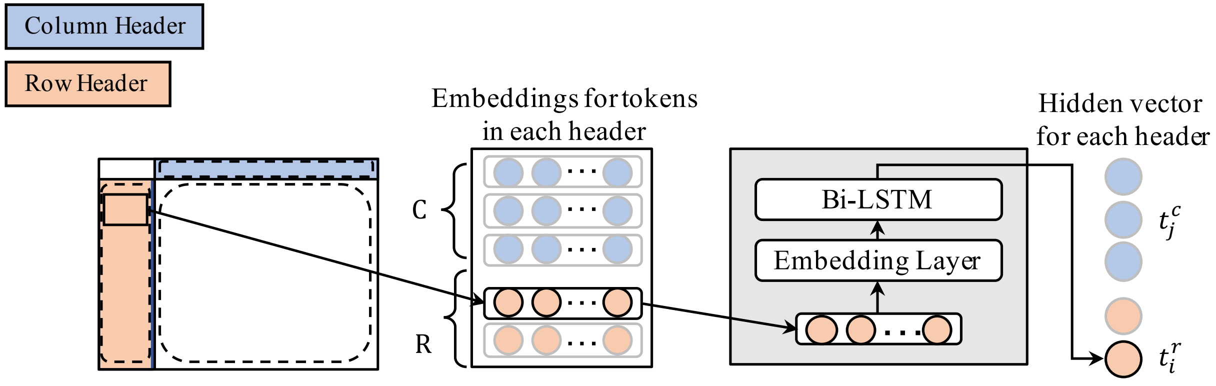
$\{\mathrm{C8}, f_{div}^{1}, \mathrm{B4}\}, \{\mathrm{C4}, f_{div}^{2}, \mathrm{B8}\}$

Formula:

B8 = +B4+B5+B6+B7

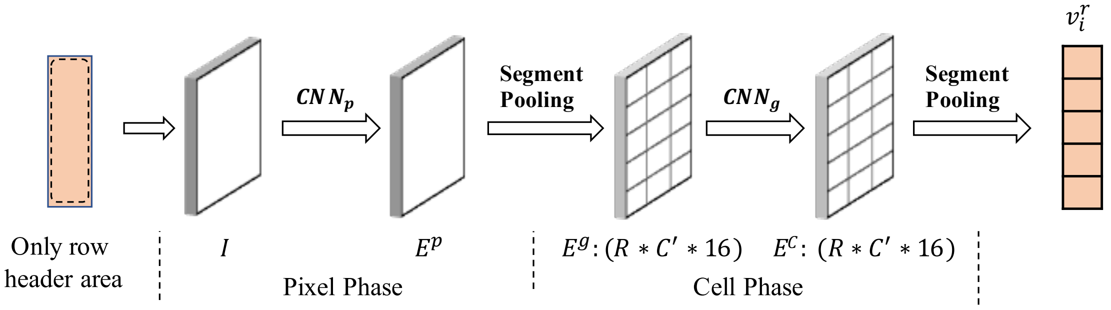
C4 = B4/B8

Solution – Cell Encoding model


Column Header
Row Header
Embeddings for tokens in each header
C
R
Bi-LSTM
Embedding Layer
Hidden vector for each header
t_j^c
t_i^r

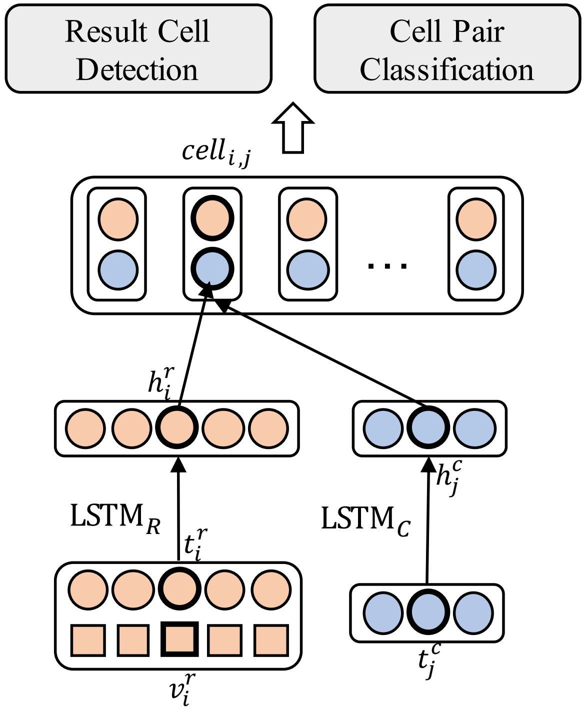
(a) Text module

Solution – Cell Encoding model



(b) Vision module

Solution – Cell Encoding model



(c) Combination and classification

Experiments

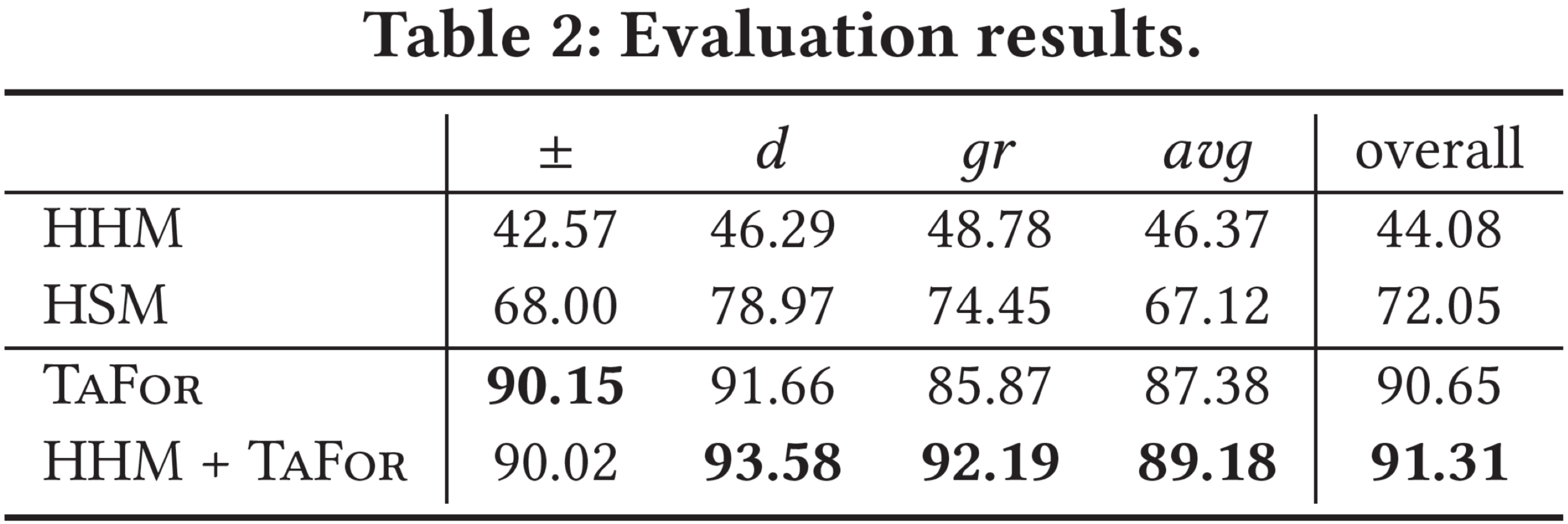


Table 2: Evaluation results.

	±	*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**

Experiments

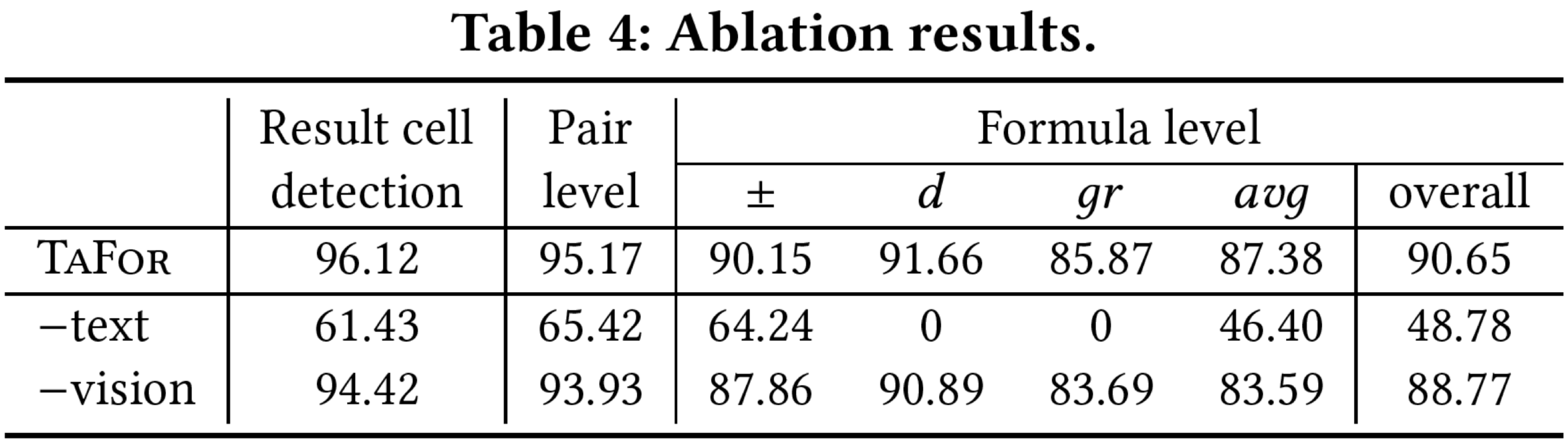
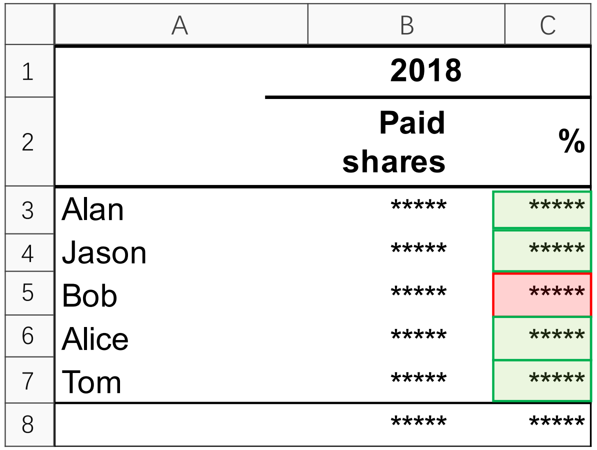


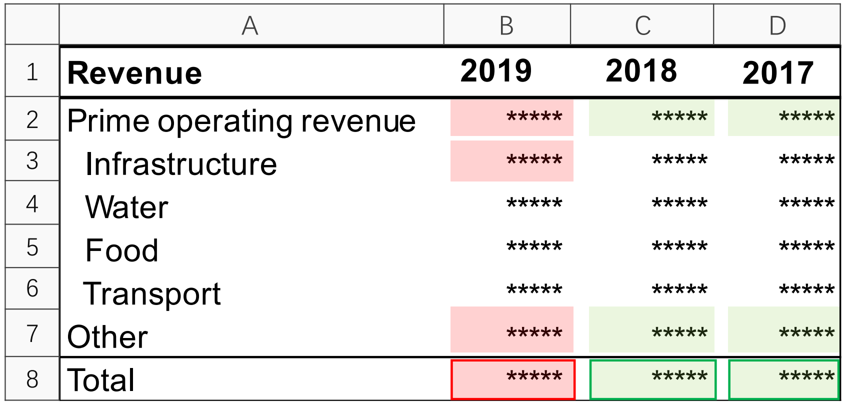
Table 4: Ablation results.

	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

Bad Cases



	A	B	C
1		2018	
2		Paid shares	%
3	Alan	*****	*****
4	Jason	*****	*****
5	Bob	*****	*****
6	Alice	*****	*****
7	Tom	*****	*****
8		*****	*****



	A	B	C	D
1	**Revenue**	**2019**	**2018**	**2017**
2	Prime operating revenue	*****	*****	*****
3	Infrastructure	*****	*****	*****
4	Water	*****	*****	*****
5	Food	*****	*****	*****
6	Transport	*****	*****	*****
7	Other	*****	*****	*****
8	Total	*****	*****	*****

Future Work

- Named entity recognition in tables.
- Consider the common sense and prior knowledge.
- Combine deep learning and symbolic knowledge.

THANK YOU