Zero-shot Key Information Extraction from Mixed-Style Tables: Pre-training on Wikipedia

Qingping Yang[1,2], Yingpeng Hu[1,2], Rongyu Cao[1,2], Hongwei Li[3], Ping Luo[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.

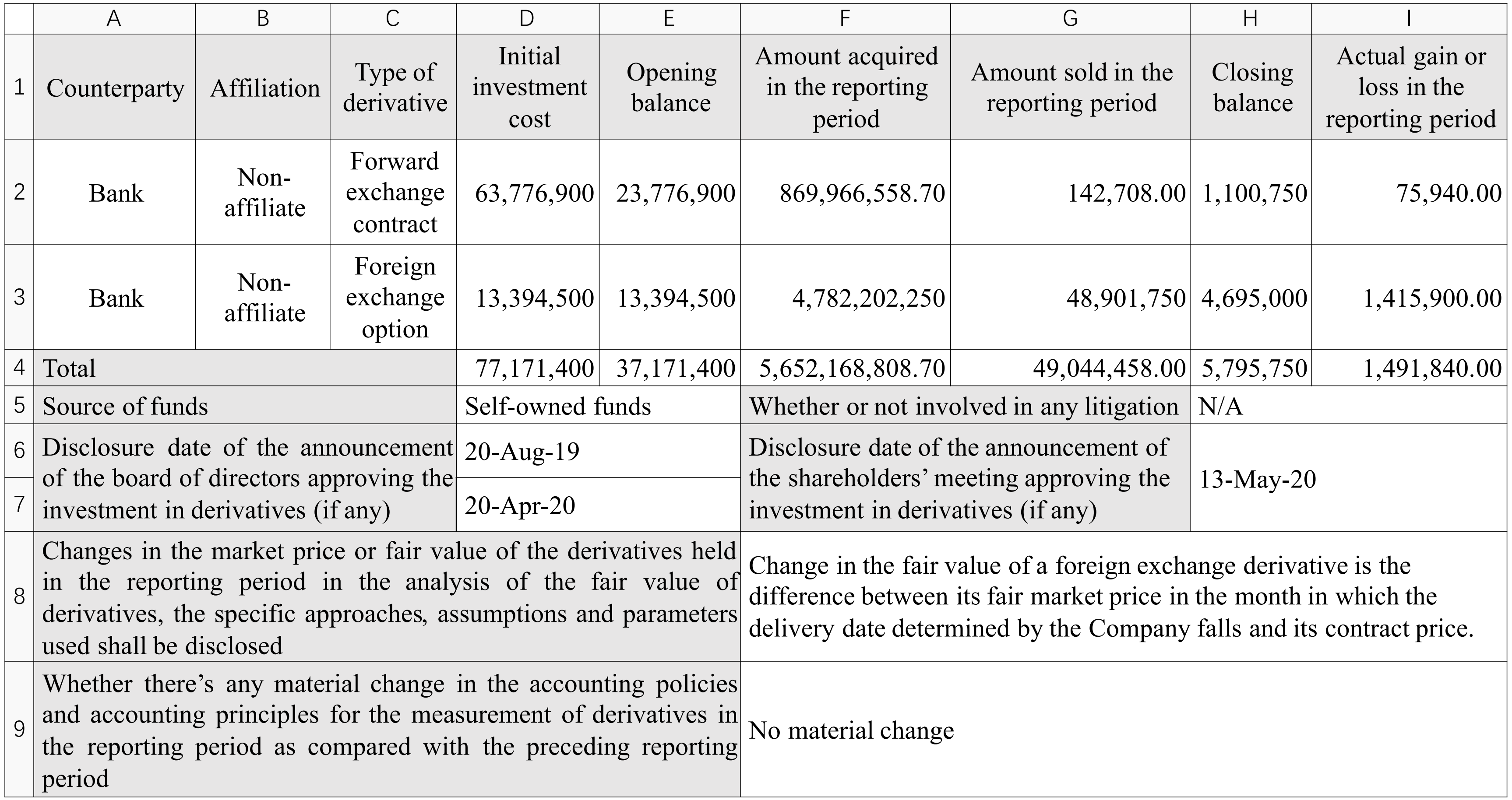
[4] Peng Cheng Laboratory

Background

- Table, an intuitive and easy-to-use tool for efficiently organizing, presenting a collection of facts, is widely used on the Web and in enterprises.

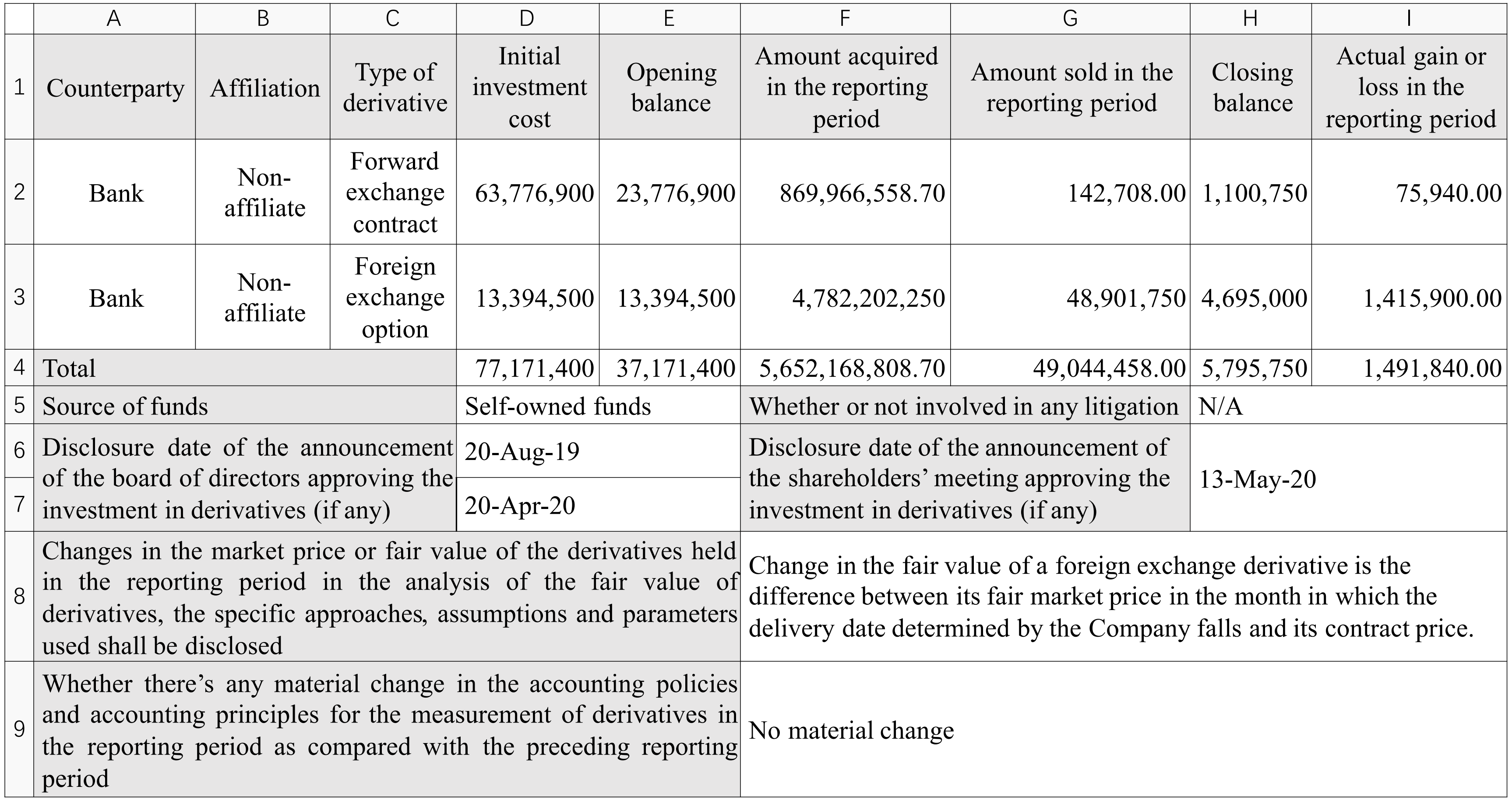
- There is always strong demand to **extract key information from tables** for further analysis.

Background



	A	B	C	D	E	F	G	H	I
1	Counterparty	Affiliation	Type of derivative	Initial investment cost	Opening balance	Amount acquired in the reporting period	Amount sold in the reporting period	Closing balance	Actual gain or loss in the reporting period
2	Bank	Non-affiliate	Forward exchange contract	63,776,900	23,776,900	869,966,558.70	142,708.00	1,100,750	75,940.00
3	Bank	Non-affiliate	Foreign exchange option	13,394,500	13,394,500	4,782,202,250	48,901,750	4,695,000	1,415,900.00
4	Total			77,171,400	37,171,400	5,652,168,808.70	49,044,458.00	5,795,750	1,491,840.00
5	Source of funds			Self-owned funds		Whether or not involved in any litigation		N/A	
6	Disclosure date of the announcement of the board of directors approving the investment in derivatives (if any)			20-Aug-19		Disclosure date of the announcement of the shareholders' meeting approving the investment in derivatives (if any)		13-May-20	
7				20-Apr-20					
8	Changes in the market price or fair value of the derivatives held in the reporting period in the analysis of the fair value of derivatives, the specific approaches, assumptions and parameters used shall be disclosed					Change in the fair value of a foreign exchange derivative is the difference between its fair market price in the month in which the delivery date determined by the Company falls and its contract price.			
9	Whether there's any material change in the accounting policies and accounting principles for the measurement of derivatives in the reporting period as compared with the preceding reporting period					No material change			

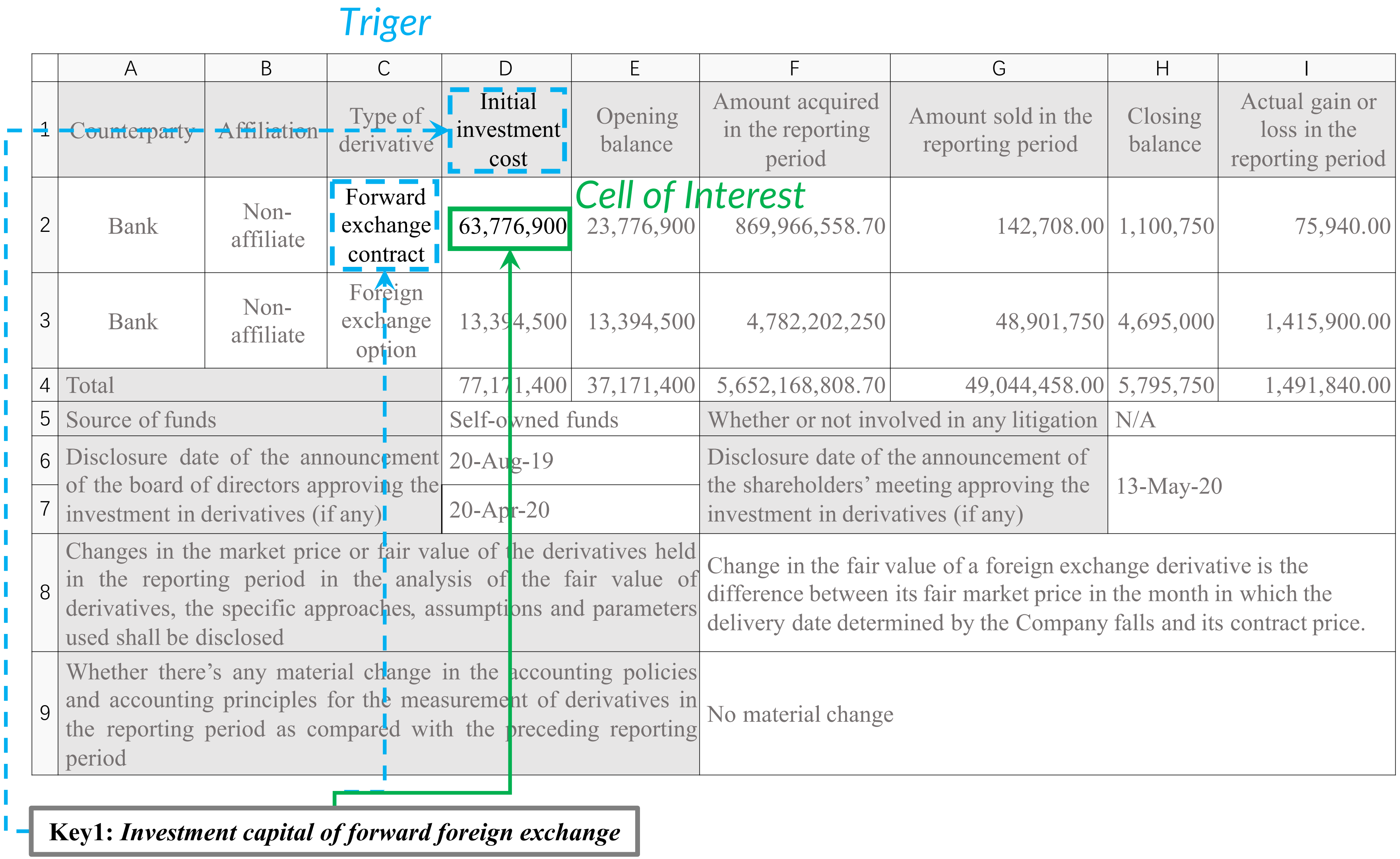
Background



	A	B	C	D	E	F	G	H	I
1	Counterparty	Affiliation	Type of derivative	Initial investment cost	Opening balance	Amount acquired in the reporting period	Amount sold in the reporting period	Closing balance	Actual gain or loss in the reporting period
2	Bank	Non-affiliate	Forward exchange contract	63,776,900	23,776,900	869,966,558.70	142,708.00	1,100,750	75,940.00
3	Bank	Non-affiliate	Foreign exchange option	13,394,500	13,394,500	4,782,202,250	48,901,750	4,695,000	1,415,900.00
4	Total			77,171,400	37,171,400	5,652,168,808.70	49,044,458.00	5,795,750	1,491,840.00
5	Source of funds			Self-owned funds		Whether or not involved in any litigation		N/A	
6	Disclosure date of the announcement of the board of directors approving the investment in derivatives (if any)			20-Aug-19		Disclosure date of the announcement of the shareholders' meeting approving the investment in derivatives (if any)		13-May-20	
7				20-Apr-20					
8	Changes in the market price or fair value of the derivatives held in the reporting period in the analysis of the fair value of derivatives, the specific approaches, assumptions and parameters used shall be disclosed					Change in the fair value of a foreign exchange derivative is the difference between its fair market price in the month in which the delivery date determined by the Company falls and its contract price.			
9	Whether there's any material change in the accounting policies and accounting principles for the measurement of derivatives in the reporting period as compared with the preceding reporting period					No material change			

Key1: *Investment capital of forward foreign exchange*

Background



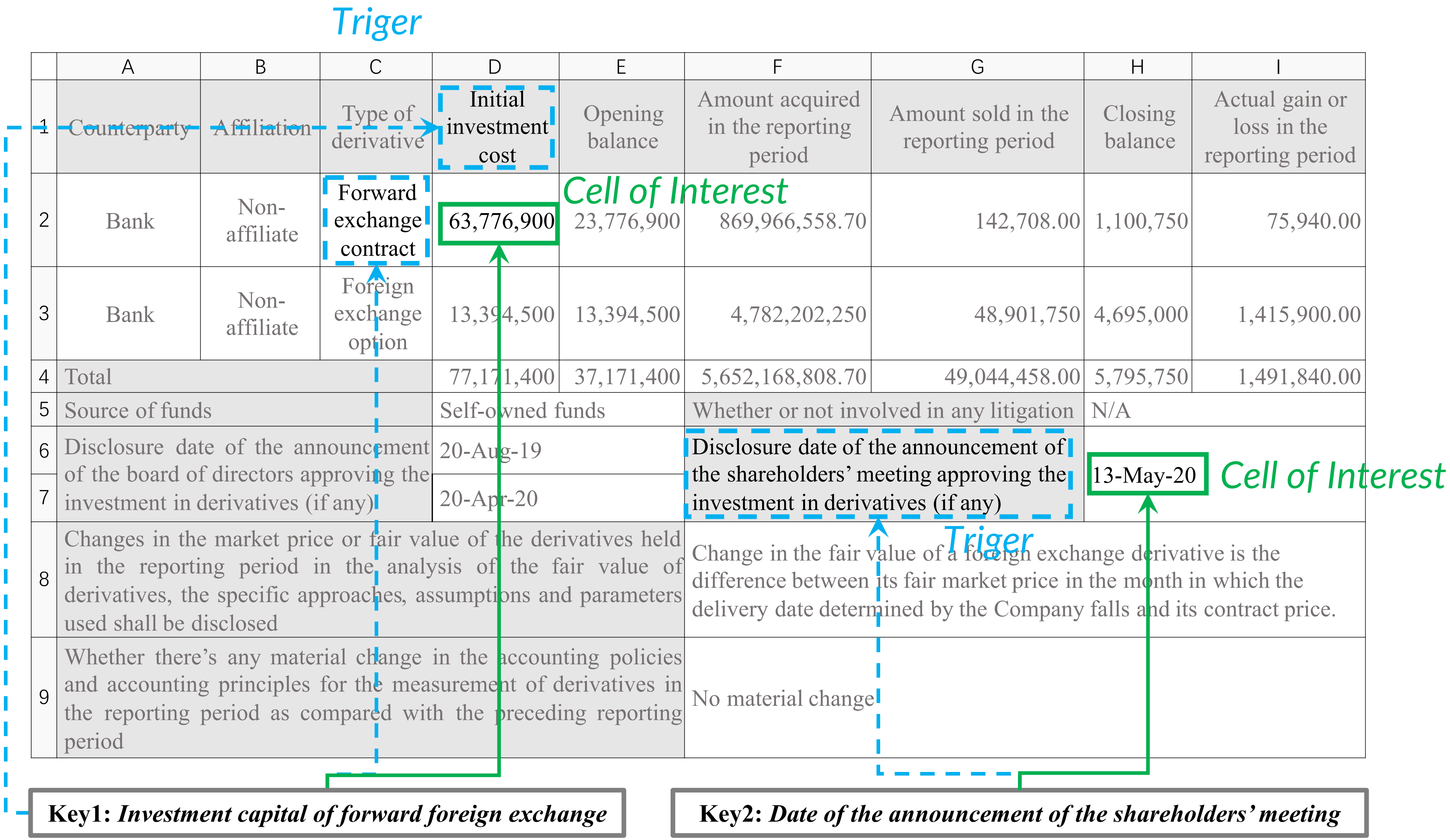
Triger

Cell of Interest

Key1: ***Investment capital of forward foreign exchange***

	A	B	C	D	E	F	G	H	I
1	Counterparty	Affiliation	Type of derivative	Initial investment cost	Opening balance	Amount acquired in the reporting period	Amount sold in the reporting period	Closing balance	Actual gain or loss in the reporting period
2	Bank	Non-affiliate	Forward exchange contract	63,776,900	23,776,900	869,966,558.70	142,708.00	1,100,750	75,940.00
3	Bank	Non-affiliate	Foreign exchange option	13,394,500	13,394,500	4,782,202,250	48,901,750	4,695,000	1,415,900.00
4	Total			77,171,400	37,171,400	5,652,168,808.70	49,044,458.00	5,795,750	1,491,840.00
5	Source of funds			Self-owned funds		Whether or not involved in any litigation		N/A	
6	Disclosure date of the announcement of the board of directors approving the investment in derivatives (if any)			20-Aug-19		Disclosure date of the announcement of the shareholders' meeting approving the investment in derivatives (if any)		13-May-20	
7				20-Apr-20					
8	Changes in the market price or fair value of the derivatives held in the reporting period in the analysis of the fair value of derivatives, the specific approaches, assumptions and parameters used shall be disclosed					Change in the fair value of a foreign exchange derivative is the difference between its fair market price in the month in which the delivery date determined by the Company falls and its contract price.			
9	Whether there's any material change in the accounting policies and accounting principles for the measurement of derivatives in the reporting period as compared with the preceding reporting period					No material change			

Background



Triger

	A	B	C	D	E	F	G	H	I
1	Counterparty	Affiliation	Type of derivative	Initial investment cost	Opening balance	Amount acquired in the reporting period	Amount sold in the reporting period	Closing balance	Actual gain or loss in the reporting period
2	Bank	Non-affiliate	Forward exchange contract	63,776,900	23,776,900	869,966,558.70	142,708.00	1,100,750	75,940.00
3	Bank	Non-affiliate	Foreign exchange option	13,394,500	13,394,500	4,782,202,250	48,901,750	4,695,000	1,415,900.00
4	Total			77,171,400	37,171,400	5,652,168,808.70	49,044,458.00	5,795,750	1,491,840.00
5	Source of funds			Self-owned funds		Whether or not involved in any litigation		N/A	
6	Disclosure date of the announcement of the board of directors approving the investment in derivatives (if any)			20-Aug-19		Disclosure date of the announcement of the shareholders' meeting approving the investment in derivatives (if any)		13-May-20	
7				20-Apr-20					
8	Changes in the market price or fair value of the derivatives held in the reporting period in the analysis of the fair value of derivatives, the specific approaches, assumptions and parameters used shall be disclosed					Change in the fair value of a foreign exchange derivative is the difference between its fair market price in the month in which the delivery date determined by the Company falls and its contract price.			
9	Whether there's any material change in the accounting policies and accounting principles for the measurement of derivatives in the reporting period as compared with the preceding reporting period					No material change			

Cell of Interest

Cell of Interest

Triger

Key1: ***Investment capital of forward foreign exchange***

Key2: ***Date of the announcement of the shareholders' meeting***

Task

Key Information Extraction (KIE) from tables.

- Taking a table and a key as input (without triggers),
- Outputting a cell from table containing the corresponding value, which output cell is called *Cell of Interest (CoI)*.

Related Work

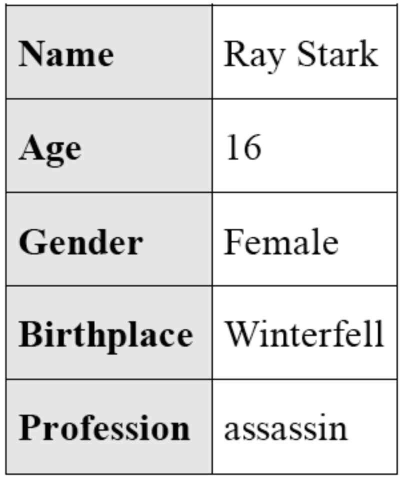
- KIE from invoices or receipts[1,2]
  1. Invoices or receipts are presented in the form of images.
  2. Only single-digit keys need to extract (e.g. 4 fields in SROIE).
  3. Cannot cover keys/fields that the model has not seen.

[1] B. P. Majumder, N. Potti, S. Tata, J. B. Wendt, Q. Zhao, and M. Najork,"Representation learning for information extraction from form-like documents," in ACL, 2020.

[2] R. Cao and P. Luo, "Extracting zero-shot structured information from form-like documents: Pretraining with keys and triggers," in AAAI, 2021.

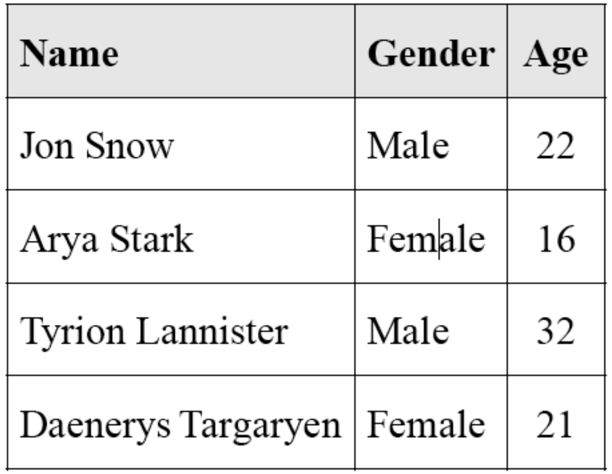
Related Work

- KIE from Tables[1]
- Question Answering on Tables[2]
  - Require relatively fixed table headers to identify table content (e.g., relational tables and entity tables).



Name	Ray Stark
Age	16
Gender	Female
Birthplace	Winterfell
Profession	assassin

Entity table



Name	**Gender**	**Age**
Jon Snow	Male	22
Arya Stark	Female	16
Tyrion Lannister	Male	32
Daenerys Targaryen	Female	21

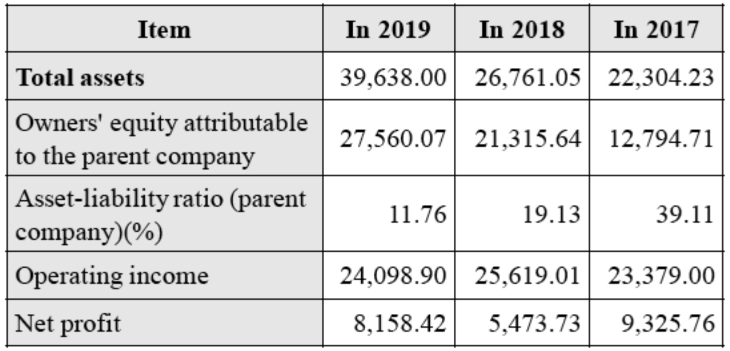
Relational table

[1] *Y. W. Wong, D. Widdows, T. Lokovic, and K. Nigam, "Scalable attribute-value extraction from semi-structured text," in ICDM, 2009.*

[2] *J. Herzig, P. K. Nowak, T. M¨uller, F. Piccinno, and J. M. Eisenschlos, "TAPAS: Weakly supervised table parsing via pre-training," in ACL, 2020.*

Related Work

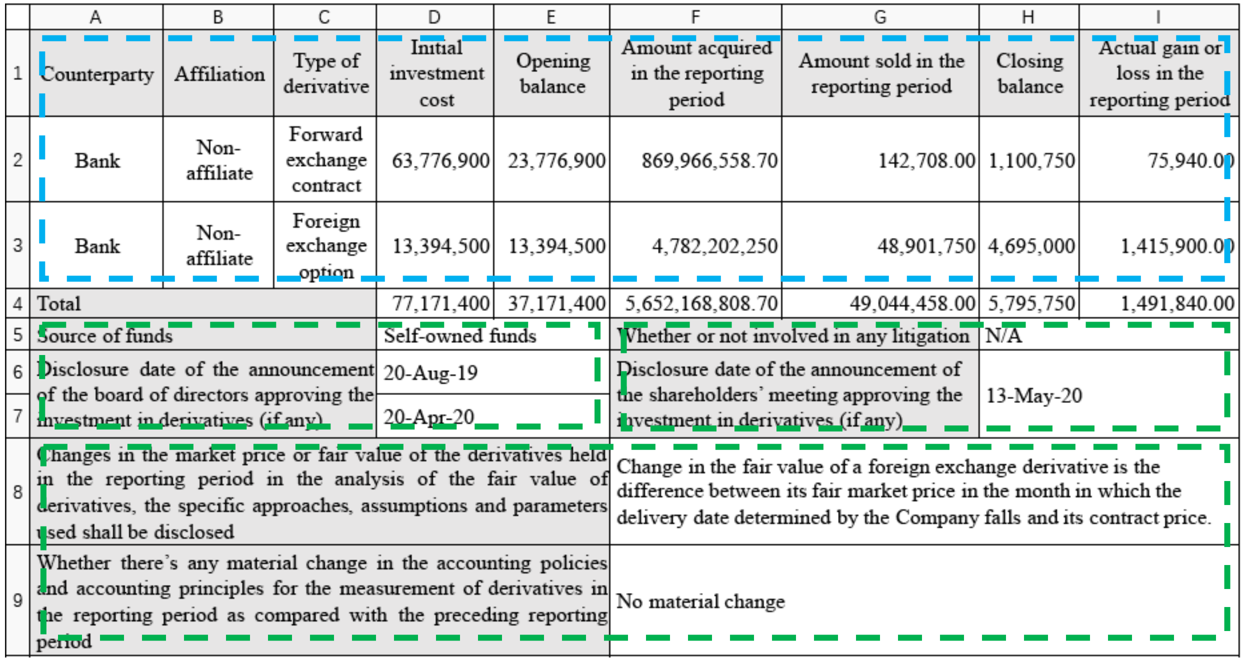
- Matrix tables and mix-style tables play a more important role especially in the financial sector.
  - In our financial dataset, the proportion of matrix tables and mixed tables are higher than 90%.



Item	In 2019	In 2018	In 2017
Total assets	39,638.00	26,761.05	22,304.23
Owners' equity attributable to the parent company	27,560.07	21,315.64	12,794.71
Asset-liability ratio (parent company)(%)	11.76	19.13	39.11
Operating income	24,098.90	25,619.01	23,379.00
Net profit	8,158.42	5,473.73	9,325.76

Matrix table

relational sub-table



	A	B	C	D	E	F	G	H	I
1	Counterparty	Affiliation	Type of derivative	Initial investment cost	Opening balance	Amount acquired in the reporting period	Amount sold in the reporting period	Closing balance	Actual gain or loss in the reporting period
2	Bank	Non-affiliate	Forward exchange contract	63,776,900	23,776,900	869,966,558.70	142,708.00	1,100,750	75,940.00
3	Bank	Non-affiliate	Foreign exchange option	13,394,500	13,394,500	4,782,202,250	48,901,750	4,695,000	1,415,900.00
4	Total			77,171,400	37,171,400	5,652,168,808.70	49,044,458.00	5,795,750	1,491,840.00
5	Source of funds			Self-owned funds		Whether or not involved in any litigation		N/A	
6	Disclosure date of the announcement of the board of directors approving the investment in derivatives (if any)			20-Aug-19		Disclosure date of the announcement of the shareholders' meeting approving the investment in derivatives (if any)		13-May-20	
7				20-Apr-20					
8	Changes in the market price or fair value of the derivatives held in the reporting period in the analysis of the fair value of derivatives, the specific approaches, assumptions and parameters used shall be disclosed					Change in the fair value of a foreign exchange derivative is the difference between its fair market price in the month in which the delivery date determined by the Company falls and its contract price.			
9	Whether there's any material change in the accounting policies and accounting principles for the measurement of derivatives in the reporting period as compared with the preceding reporting period					No material change			

entity sub-table

Mix-style table

Contributions

Our main contributions:

1. To the best of our knowledge, this paper is the first work to explore KIE from mixed-style tables.
2. Our model captures the semantics of keys to address the issue of zero-shot keys.
3. The experiments on a financial dataset show that the proposed model is effective, and obtains great improvement in accuracy on zero-shot keys with the pre-training.

Problem Formulation

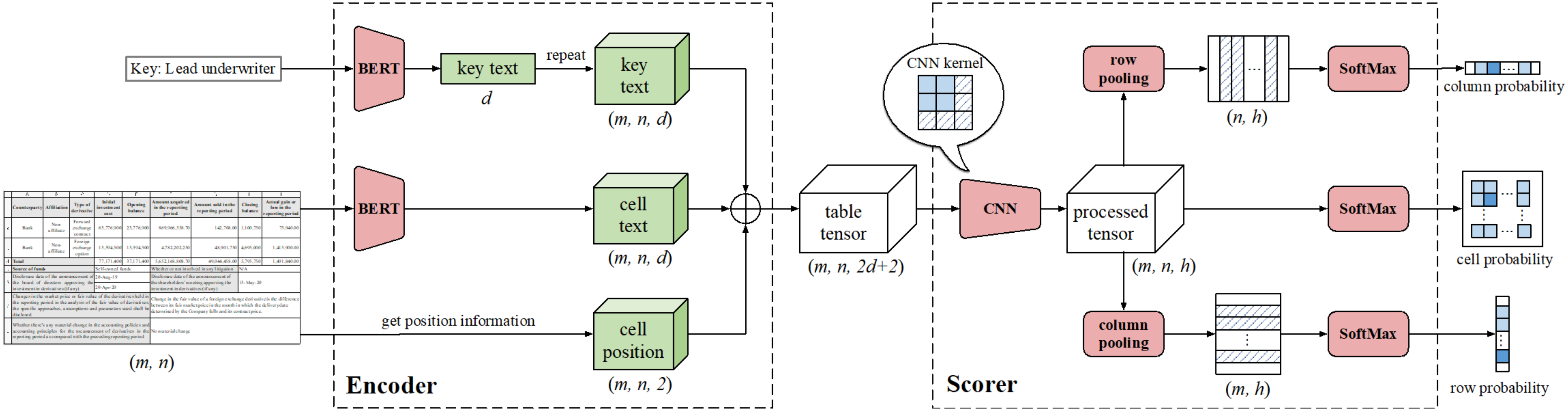
Zero-shot keys: $\mathcal{K}_n = \{k_i\}_{i=1}^{N_n}$, non-zero-shot keys: $\mathcal{K}_z = \{k_i\}_{i=1}^{N_z}$

Training set: $D_{tr} = \{(k_i, T_i, c_i^*) | k_i \in \mathcal{K}_n\}$

Test set: $D_{te} = \{(k_i, T_i, c_i^*) | k_i \in \mathcal{K}_n \cup \mathcal{K}_z\}$

The probability of being the CoI of the cell $c_{\langle i,j \rangle}$: $P(c_{\langle i,j \rangle} | k_i, T_i)$

IEMT: Information Extraction from Mixed-style Table



Counterparty	Affiliation	Type of derivative	Initial investment cost	Opening balance	Amount acquired in the reporting period	Amount sold in the reporting period	Closing balance	Actual gain or loss in the reporting period
Bank	Non-affiliate	Forward exchange contract	63,776,900	23,776,900	869,966,558.70	142,708.00	1,100,750	75,940.00
Bank	Non-affiliate	Foreign exchange option	13,394,500	13,394,500	4,782,202,250	48,901,750	4,695,000	1,415,900.00
Total			77,171,400	37,171,400	5,652,168,808.70	49,044,458.00	5,795,750	1,491,840.00
Source of funds			Self-owned funds		Whether or not involved in any litigation		N/A	
Disclosure date of the announcement of the board of directors approving the investment in derivatives(if any)			20-Aug-19 20-Apr-20		Disclosure date of the announcement of theshareholders' meeting approving the investment in derivatives(if any)		13-May-20	
Changes in the market price or fair value of the derivatives held in the reporting period in the analysis of the fair value of derivatives, the specific approaches, assumptions and parameters used shall be disclosed					Change in the fair value of a foreign exchange derivative is the difference between its fair market price in the month in which the deliverydate determined by the Company falls and its contract price.			
Whether there's any material change in the accounting policies and accounting principles for the measurement of derivatives in the reporting period as compared with the preceding reporting period					No material change			

Joint Training

1. Cell classification

$$L_{cell} = -\sum_{c \in T} [l^c \log(P(c)) + (1 - l^c) \log(1 - P(c))]$$

2. Row classification

$$L_{row} = -\sum_{i=1}^{n} [l_i^r \log(P(r_i)) + (1 - l_i^r) \log(1 - P(r_i))]$$

3. Col classification

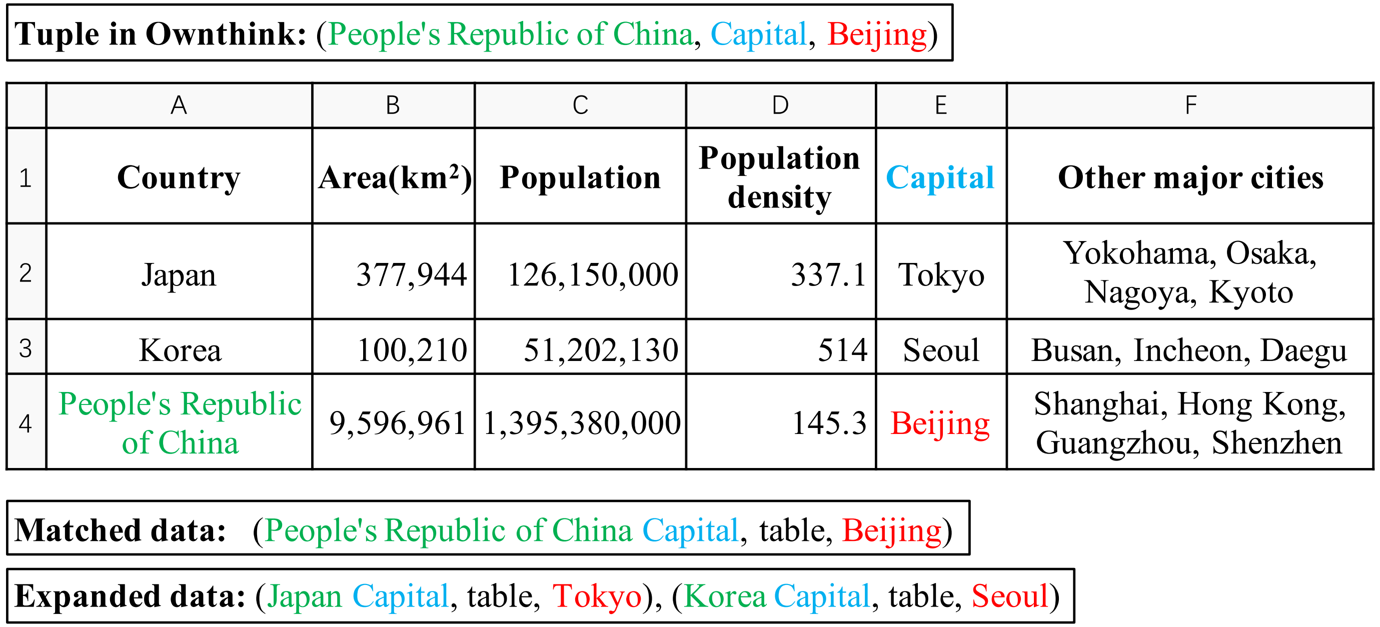
   L_{col} is calculated similar to L_{row}

Final loss function:

$$L = L_{cell} + \alpha(L_{row} + L_{col})$$

Pretraining

- Pretraining dataset:
  - Ownthink, a huge Chinese knowledge graph that contains about 140 million tuples.
  - Tables on Chinese Wikipedia.
    - We match entity, attribute and value in Ownthink with tables from Chinese Wikipedia to construct our pretraining dataset.



Tuple in Ownthink: (People's Republic of China, Capital, Beijing)

	A	B	C	D	E	F
1	**Country**	**Area(km²)**	**Population**	**Population density**	**Capital**	**Other major cities**
2	Japan	377,944	126,150,000	337.1	Tokyo	Yokohama, Osaka, Nagoya, Kyoto
3	Korea	100,210	51,202,130	514	Seoul	Busan, Incheon, Daegu
4	People's Republic of China	9,596,961	1,395,380,000	145.3	Beijing	Shanghai, Hong Kong, Guangzhou, Shenzhen

Matched data: (People's Republic of China Capital, table, Beijing)

Expanded data: (Japan Capital, table, Tokyo), (Korea Capital, table, Seoul)

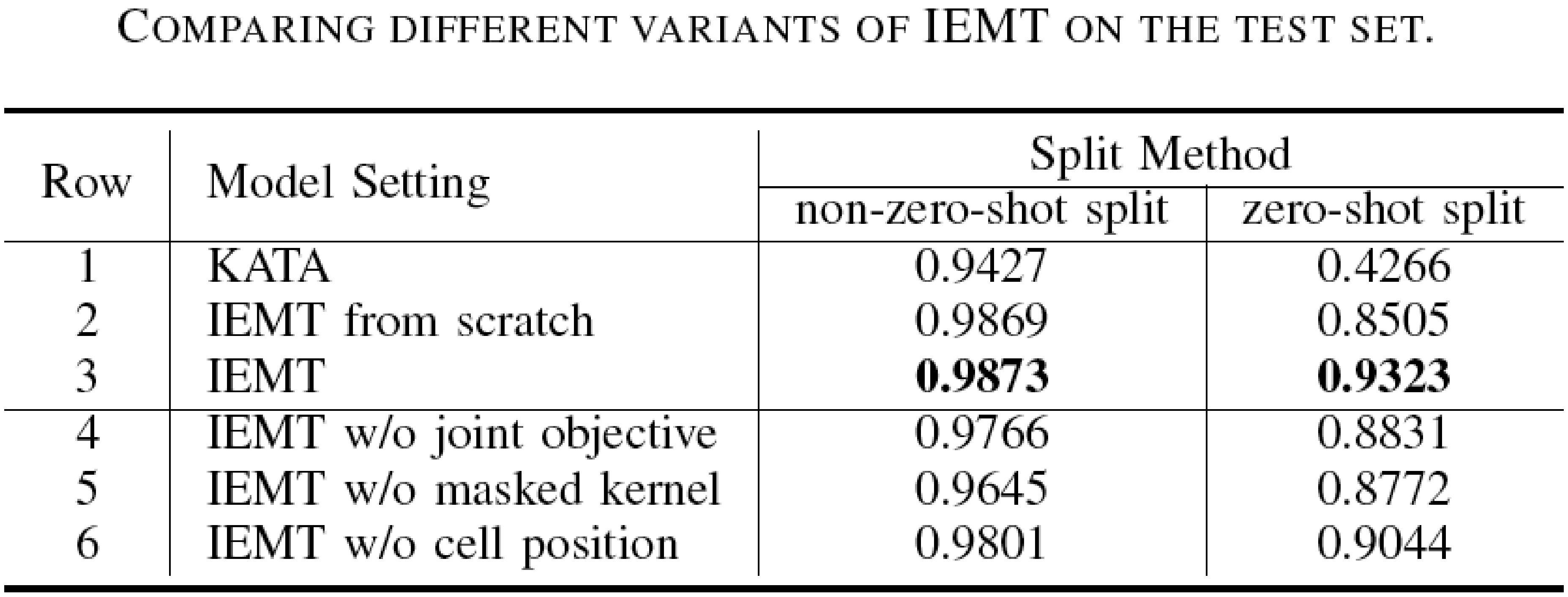
Experiments

- Baseline
  - KATA[1], which aims to extract key information from document pages, is extended by LayoutLM[2] with explicitly trigger-supervised training.

- Dataset
  - 26,869 Financial tables from CNINFO

[1] R. Cao and P. Luo, "Extracting zero-shot structured information from form-like documents: Pretraining with keys and triggers," in AAAI, 2021.
[2] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in KDD, 2020.

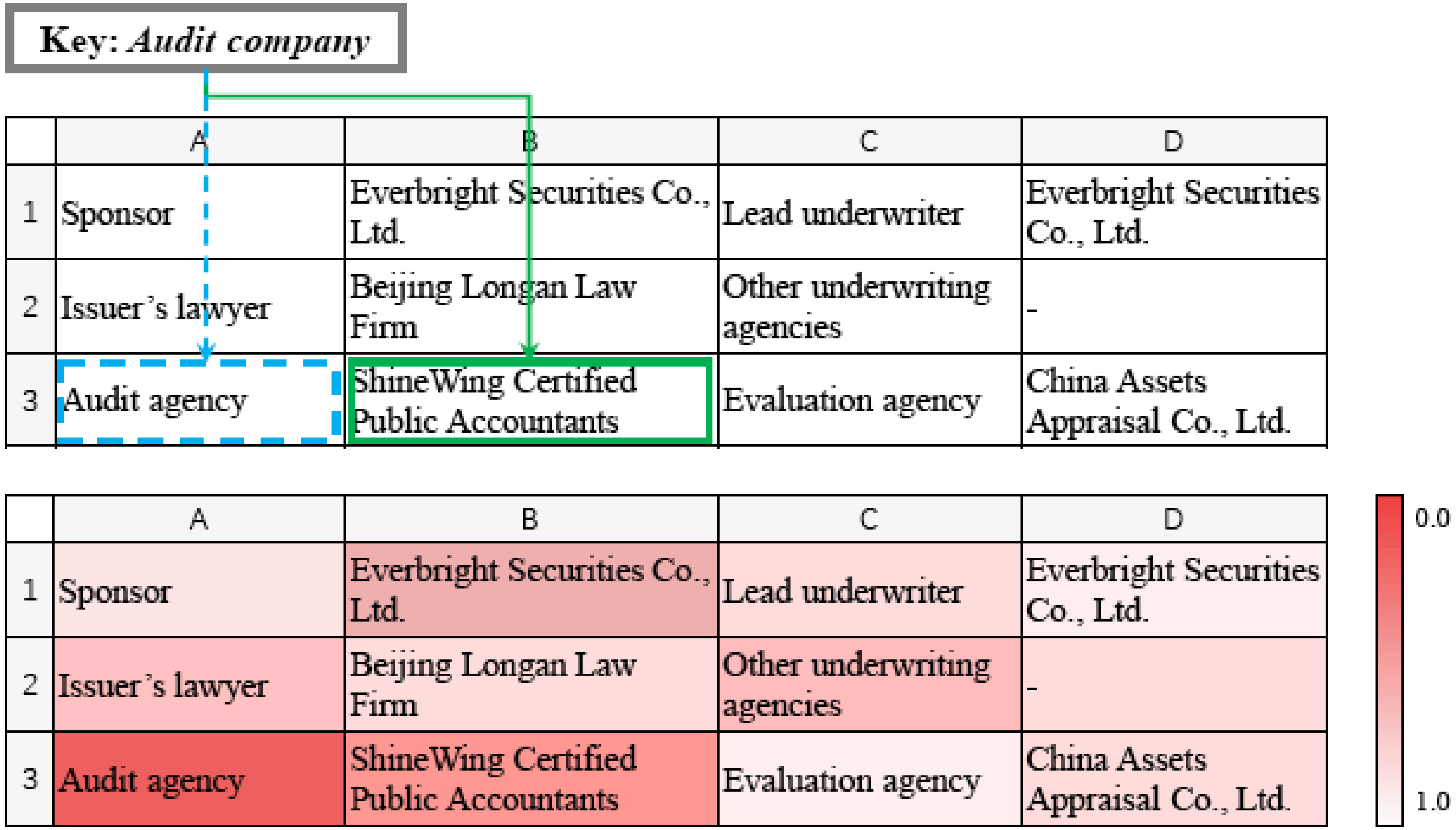
Experiments



Comparing different variants of IEMT on the test set.

Row	Model Setting	Split Method	
		non-zero-shot split	zero-shot split
1	KATA	0.9427	0.4266
2	IEMT from scratch	0.9869	0.8505
3	IEMT	**0.9873**	**0.9323**
4	IEMT w/o joint objective	0.9766	0.8831
5	IEMT w/o masked kernel	0.9645	0.8772
6	IEMT w/o cell position	0.9801	0.9044

Experiments



Key: *Audit company*

	A	B	C	D
1	Sponsor	Everbright Securities Co., Ltd.	Lead underwriter	Everbright Securities Co., Ltd.
2	Issuer's lawyer	Beijing Longan Law Firm	Other underwriting agencies	-
3	Audit agency	ShineWing Certified Public Accountants	Evaluation agency	China Assets Appraisal Co., Ltd.

	A	B	C	D
1	Sponsor	Everbright Securities Co., Ltd.	Lead underwriter	Everbright Securities Co., Ltd.
2	Issuer's lawyer	Beijing Longan Law Firm	Other underwriting agencies	-
3	Audit agency	ShineWing Certified Public Accountants	Evaluation agency	China Assets Appraisal Co., Ltd.

Fig. 5. An example to show the importance of each cell.

THANK YOU