Doc2EDAG: An End-to-End Document-level Framework for Chinese Financial Event Extraction

Shun Zheng[†], Wei Cao[‡], Wei Xu[†] and Jiang Bian[‡]

[†]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[‡]Microsoft Research Asia, Beijing, China

Motivations

- **Sentence-level Event Extraction (SEE):** merely extracting event arguments within the sentence scope (most existing research).
- **Document-level Event Extraction (DEE):** extracting events whose arguments scatter across multiple sentences of a document (practical demands).

Entity Mark Table

Mark	Entity	Entity (English)
[PER]	刘维群	Weiqun Liu
[ORG]	国信证券股份有限公司	Guosen Securities Co., ltd.
[DATE1]	2017年9月22日	Sept. 22nd, 2017
[DATE2]	2018年9月6日	Sept. 6th, 2018
[DATE3]	2018年9月20日	Sept. 20th, 2018
[DATE4]	2019年3月20日	Mar. 20th, 2019
[SHARE1]	750000股	750000 shares
[SHARE2]	975000股	975000 shares
[SHARE3]	525000股	525000 shares
[SHARE4]	1500000股	1500000 shares
[SHARE5]	16768903股	16768903 shares
[RATIO]	1. 0858%	1.0858%

Event Role / Event Record / Event Argument / Entity Mention

Event Table of Equity Pledge

Pledger	Pledged Shares	Pledgee	Begin Date	End Date	Total Holding Shares	Total Holding Ratio
[PER]	[SHARE2]	[ORG]	[DATE1]	[DATE4]	[SHARE5]	[RATIO]
[PER]	[SHARE3]	[ORG]	[DATE2]	[DATE4]	[SHARE5]	[RATIO]

ID	Sentence
5	**[DATE1]**，**[PER]**将其持有的公司**[SHARE1]**股份质押给**[ORG]**。
	In **[DATE1]**, **[PER]** pledged his **[SHARE1]** to **[ORG]**.
7	公司实施资本公积金转增股本后，其质押股份变为**[SHARE2]**。
	After the company carried out the transferring of the capital accumulation fund to the capital stock, his pledged shares became **[SHARE2]**.
8	**[DATE2]**，**[PER]**将其持有的**[SHARE3]**公司股份质押给**[ORG]**，作为对上述质押股份的补充质押。
	In **[DATE2]**, **[PER]** pledged **[SHARE3]** to **[ORG]**, as a supplementary pledge to the above pledged shares.
9	上述质押及补充质押股份合计为**[SHARE4]**，原定购回日期为**[DATE3]**。
	The aforementioned pledged and supplementary pledged shares added up to **[SHARE4]**, and the original repurchase date was **[DATE3]**.
10	**[DATE3]**，**[PER]**针对其质押的**[SHARE4]**股份办理了延期购回业务，购回日期延长至**[DATE4]**。
	In **[DATE3]**, **[PER]** extended the repurchase date to **[DATE4]** for **[SHARE4]** he pledged.
12	截至本公告日，**[PER]**持有公司股份**[SHARE5]**，占公司总股本的**[RATIO]**。
	As of the date of this announcement, **[PER]** hold **[SHARE5]** of the company, accounting for **[RATIO]** of the total share capital of the company.

Figure 1:A document with two *Equity Pledge* events to illustrate the *arguments-scattering* and *multi-event* challenges.

Contributions

- Propose the first end-to-end modeling framework for DEE.
- Formalize a novel task for DEE without specifying explicit trigger words, which can make the labeling process dramatically easier.
- Build a large-scale and high-quality dataset based on real-world applications.
- Conduct extensive experiments to demonstrate the superiority of Doc2EDAG and reveal the specific challenges of DEE.
- Open both data and codes at `https://github.com/dolphin-zs/Doc2EDAG` to facilitate future research about DEE.

Key Innovations

A Novel Task for DEE:

- **Entity Extraction**: extracting entities as candidates for event arguments.
- **Event Triggering Classification**: judging whether a document triggers a specific event type.
- **Event Table Filling**: populating the table of triggered events with extracted entities.

Entity-based directed acyclic graph (EDAG): transforming the hard table-filling task into several path-expanding sub-tasks that are more tractable.

Doc2EDAG

End-to-End Modeling for DEE:
Input Representation + Entity Recognition + Document-level Entity Encoding + EDAG Generation.

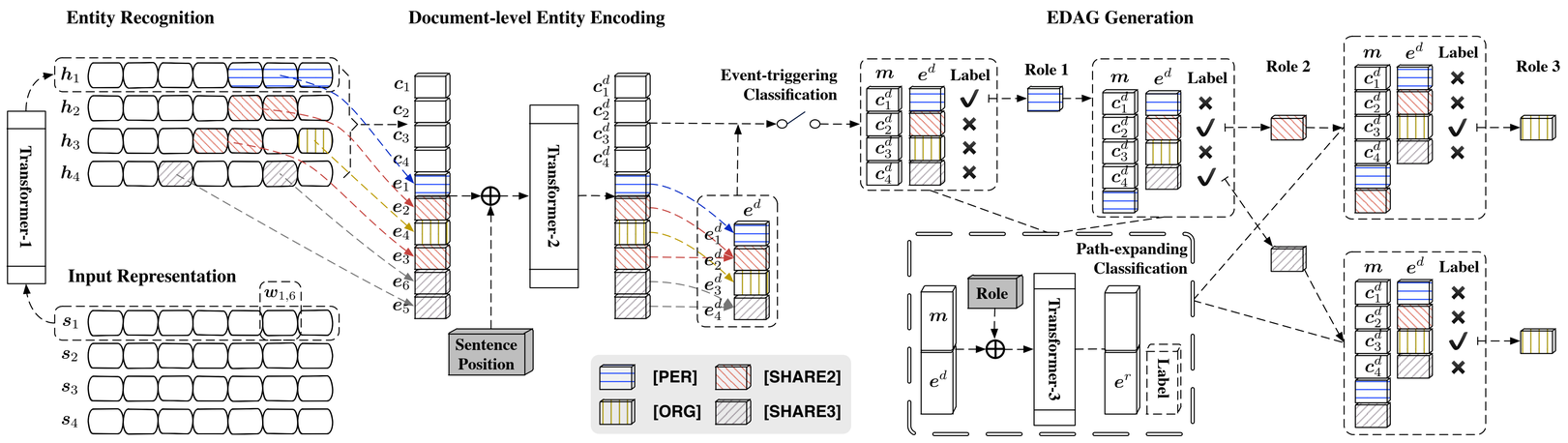

Figure 2:The overall architecture of Doc2EDAG.

Dataset

- **Data Collection:** conducting distant supervision on Chinese financial announcements (2008 - 2018) and achieving 94% F1 on 100 human-annotated documents.
- **Event Types:** *Equity Freeze* (EF), *Equity Repurchase* (ER), *Equity Underweight* (EU), *Equity Overweight* (EO), and *Equity Pledge* (EP).

Event	#Train	#Dev	#Test	#Total	MER (%)
EF	806	186	204	1,196	32.0
ER	1,862	297	282	3,677	16.1
EU	5,268	677	346	5,847	24.3
EO	5,101	570	1,138	6,017	28.0
EP	12,857	1,491	1,254	15,602	35.4
All	25,632	3,204	3,204	32,040	29.0

Table 1:Dataset statistics about the number and the multi-event ratio (MER).

Baselines

Two Variants of DCFEE (Yang et al, 2018), which uses a sequence tagging model for SEE and heuristics for DEE.

- **DCFEE-O**: one sentence one event record.
- **DCFEE-M**: one sentence multi records (just guessing).

GreedyDec: greedily generating one event record.

Experiments

- **Main Results:** Table 2.
- **Single-event vs. Multi-event:** Table 3.
- **Ablation Tests:** Table 4.

Model	EF P.	EF R.	EF F1	ER P.	ER R.	ER F1	EU P.	EU R.	EU F1	EO P.	EO R.	EO F1	EP P.	EP R.	EP F1
DCFEE-O	66.0	41.6	51.1	84.5	81.8	83.1	62.7	35.4	45.3	51.4	42.6	46.6	64.3	63.6	63.9
DCFEE-M	51.8	40.7	45.6	83.7	78.0	80.8	49.5	39.9	44.2	42.5	47.5	44.9	59.8	66.4	62.9
GreedyDec	**79.5**	46.8	58.9	83.3	74.9	78.9	68.7	40.8	51.2	69.7	40.6	51.3	**85.7**	48.7	62.1
Doc2EDAG	77.1	**64.5**	**70.2**	**91.3**	**83.6**	**87.3**	**80.2**	**65.0**	**71.8**	**82.1**	**69.0**	**75.0**	80.0	**74.8**	**77.3**

Table 2:Overall event-level precision (P.), recall (R.) and F1 scores on the test set.

Model	EF S.	EF M.	ER S.	ER M.	EU S.	EU M.	EO S.	EO M.	EP S.	EP M.	Avg. S.	Avg. M.	Avg. S. & M.
DCFEE-O	56.0	46.5	86.7	54.1	48.5	41.2	47.7	45.2	68.4	61.1	61.5	49.6	58.0
DCFEE-M	48.4	43.1	83.8	53.4	48.1	39.6	47.1	42.0	67.0	60.6	58.9	47.7	55.7
GreedyDec	75.9	40.8	81.7	49.8	62.2	34.6	65.7	29.4	**88.5**	42.3	74.8	39.4	60.5
Doc2EDAG	**80.0**	**61.3**	**89.4**	**68.4**	**77.4**	**64.6**	**79.4**	**69.5**	85.5	**72.5**	**82.3**	**67.3**	**76.3**

Table 3:F1 scores on the single-event (S.) and multi-event (M.) sets.

Model	EF	ER	EU	EO	EP	Avg.
Doc2EDAG	70.2	87.3	71.8	75.0	77.3	76.3
-PathMem	-11.2	-0.2	-10.1	-16.3	-10.9	-9.7
-SchSamp	-5.3	-4.8	-5.3	-6.6	-3.0	-5.0
-DocEnc	-4.7	-1.5	-1.6	-1.1	-1.5	-2.1
-NegCW	-1.4	-0.4	-0.7	-1.3	-0.4	-0.8

Table 4:F1 scores on the test set under ablation tests.