LEVEN: A Large-Scale Chinese Legal Event Detection Dataset

Feng Yao*, Chaojun Xiao*, Xiaozhi Wang, Zhiyuan Liu,
Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, Maosong Sun
Tsinghua University, Powerlaw Intelligent Technology

Overview

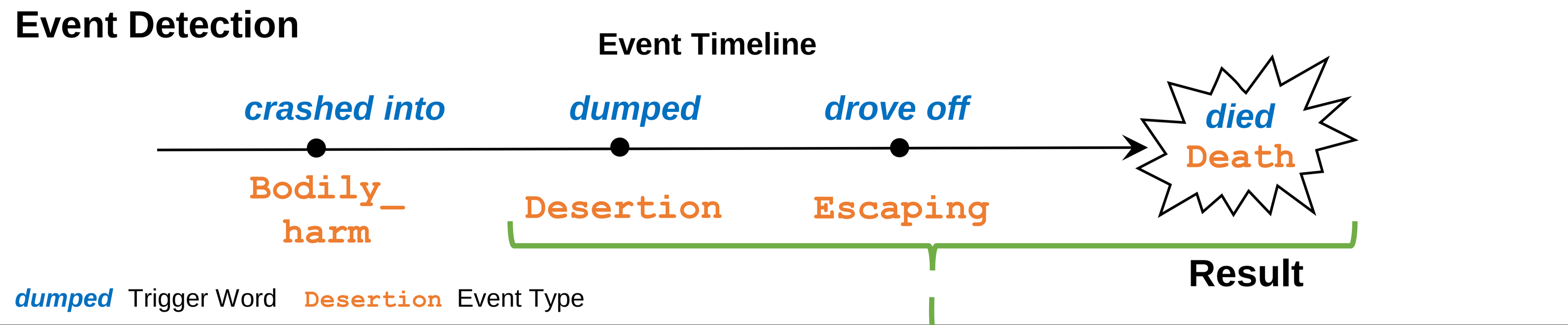
Motivation

Events are the essence of the facts in legal cases. Therefore, Legal Event Detection (LED) is fundamentally important and naturally beneficial to case understanding and other Legal AI tasks.

Fact Description

Alice drove a car at night and ***crashed into*** Bob, a pedestrian, on Green Avenue. To prevent being spotted, Alice took Bob away from the scene, ***dumped*** him under an isolated bridge and ***drove off*** in a panic. Two hours later, Bob ***died*** of excessive bleeding …

Event Detection



Related Law Article

Traffic accident crime … if the ***hit-and-run*** occurs, the crime should be sentenced to *imprisonment more than 3 years but less than 7 years* … if the perpetrator ***abandons*** the victim, resulting in the ***death***, he shall be convicted of Intentional homicide crime and sentenced to *death, life imprisonment or imprisonment of no less than 10 years* …

Crime & Prison Term

Intentional homicide crime; 10 years and 6 months

Challenges

Existing LED datasets suffer from 1) **Limited Data** and 2) **Incomprehensive Event Schema**.

Dataset	#Documents	#Tokens	#Sentences	#Event Types	#Event Mentions	Language	Domain
MAVEN	4,480	1,276k	49,873	168	118,732	English	General
ACE2005-zh	633	185k	7,955	33	4,090	Chinese	General
DuEE	11,224	530k	16,900	65	19,640	Chinese	General
DivorceEE*	3,100	–	–	13	–	Chinese	Legal
CLEE*	3,000	–	6,538	5	6,538	Chinese	Legal
DyHiLED*	–	–	–	11	2,380	Chinese	Legal
LEVEN	8,116	2,241k	63,616	108	150,977	Chinese	Legal

Features

- **Large-Scale**
  LEVEN is the **largest Legal ED** dataset and the **largest Chinese ED** dataset.
- **High-Coverage**
  LEVEN covers not only **charge-oriented** events, but also **general** events.

Top-level Event Type	Category	#Type	#Mention	Percentage	Sub-type Examples
General_behaviors	Behavior	40	68,616	45.4%	Selling, Employing
Prohibited_acts	Behavior	40	43,021	28.5%	Killing, Blackmail, Theft
Judicature_related	Behavior	13	29,709	19.7%	Arrest, Surrendering
Consequences	Result	7	6,832	4.5%	Death, Injury, Being_trapped
Accident	Result	4	2,742	1.8%	Traffic_accident, Fire_acc
Natural_disaster	Majeure	4	57	0.03%	Drought, Flood&waterlogging

Event Detection Baselines

- **LEVEN Data Splits**

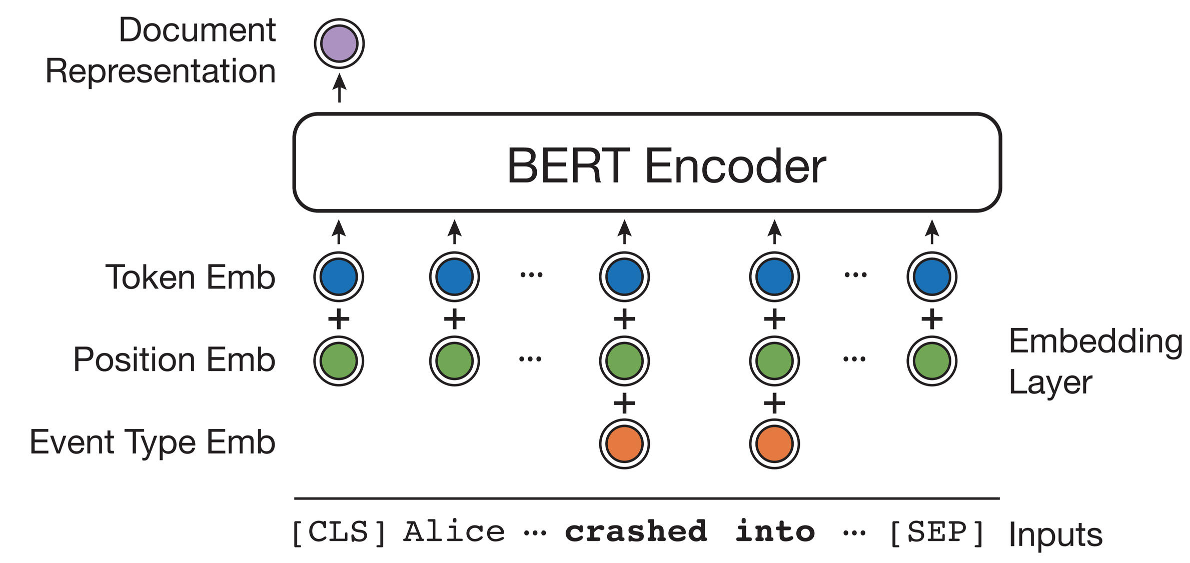
	#Documents.	#Sentences	#Event Mention	#Negative.
Training	5,301	41,238	98,410	297,252
Validation	1,230	9,788	22,885	69,645
Test	1,585	12,590	29,682	90,512

- **Test Performances of ED Baselines**

Model	Micro Precision	Micro Recall	Micro F1	Macro Precision	Macro Recall	Macro F1
DMCNN	**85.88** ± 0.70	79.70 ± 0.59	82.67 ± 0.08	80.55 ±0.49	73.31 ± 3.88	75.03 ± 0.40
BiLSTM	83.09 ± 0.89	85.16 ± 0.95	84.11 ± 0.24	78.70 ± 0.92	76.67 ± 2.23	76.65 ± 1.42
BiLST+CRF	84.74 ± 0.55	83.33 ± 0.49	84.03 ± 0.05	78.56 ± 1.31	72.60 ± 1.11	74.49 ± 0.77
BERT	84.19 ± 0.39	84.31 ± 0.34	84.25 ± 0.18	79.61 ± 0.91	76.76 ± 1.79	77.33 ± 1.30
BERT+CRF	83.82 ± 0.48	84.56 ± 0.52	84.19 ± 0.09	79.77 ± 1.10	77.65 ± 2.20	77.84 ± 1.58
DMBERT	84.77 ± 0.91	**86.22** ± 0.77	**85.48** ± 0.18	**81.57** ± 1.04	**80.90** ± 1.38	**80.34** ± 0.74

Downstream Legal AI Applications

- **Encoder Architecture**



- **Legal Judgment Prediction with event on CAIL2018**

Model	Charge Precision	Charge Rcall	Charge Mic-F1	Law Precision	Law Recall	Law Mic-F1	Term Log Distance ↓
50-shot							
BERT	76.6	**77.0**	76.8	73.6	**76.8**	75.2	2.398
+ event	**79.2**	76.2	**77.7**	**75.4**	75.6	**75.5**	**2.364**
full							
BERT	88.2	89.4	88.8	83.7	86.8	85.2	1.895
+ event	**88.2**	**89.7**	**88.9**	**83.8**	**87.7**	**85.7**	**1.878**

- **Similar Case Retrieval with event on LeCaRD**

Model	MAP	NDCG@10	NDCG@20	NDCG@30	P@5	P@10
BM25	48.40	73.10	79.70	88.80	40.60	38.10
TFIDF	45.70	79.50	83.20	84.80	30.40	26.10
LMIR	49.50	76.90	81.80	90.00	43.60	40.60
Bag-of-Event	50.94	78.37	83.66	90.32	44.11	42.62
Bag-of-Event$_w$	**51.02**	**79.90**	**84.42**	**90.97**	**45.23**	**43.36**
BERT	51.92	79.23	84.12	91.28	44.49	40.10
+ event	**51.99**	**80.10**	**84.92**	**91.73**	**44.63**	**41.22**

Performance Analysis

- **Performance of DMBERT on top-level Event Types**

Top-level Event Type	Precision	Recall	Micro-F1
General_behaviors	83.71	85.67	84.86
Prohibited_acts	83.01	82.93	82.97
Judicature_related	94.17	91.89	**93.01**
Consequences	84.54	82.92	83.73
Accident	86.04	84.40	85.21
Natural_disaster	77.78	63.64	**70.00**

- **Performance of DMBERT on Long-tail Event Types**

F1-score	[0,0.4)	[0.4,0.6)	[0.6,0.8)	[0.8,0.9)	[0.9,1.0]	sum
#low-frequency	5	4	4	**4**	**4**	21
#mid-frequency	0	0	9	13	6	28
#high-frequency	0	0	14	23	22	59

Here, low-freq and high-freq represent the number of event types that have less than 150 event mentions and more than 500 event mentions. And mid-freq denotes the remaining.

Code and Paper

https://github.com/thunlp/LEVEN

https://arxiv.org/abs/2203.08556