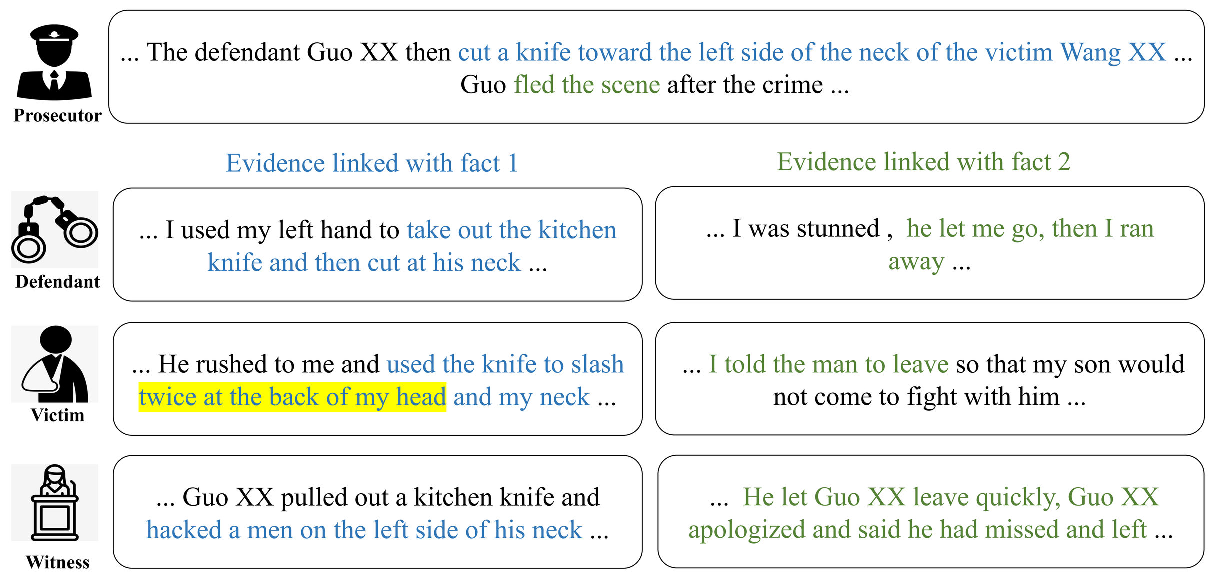
Unsupervised Legal Evidence Retrieval via Contrastive Learning with Approximate Aggregated Positive

Feng Yao, Jingyuan Zhang, Yating Zhang, Xiaozhong Liu,
Changlong Sun, Yun Liu, Weixing Shen
Tsinghua University, Alibaba DAMO Academy

Background

Motivation

Each fact requires relevant evidence to support. Before the trial, judges need to review the facts submitted by prosecutors via searching for relevant evidence within the same case, which is tedious and time-consuming. Thus, we propose the Legal Evidence Retrieval (LER) task which aims to automatically retrieve relevant evidence for each fact.



Dataset Construction

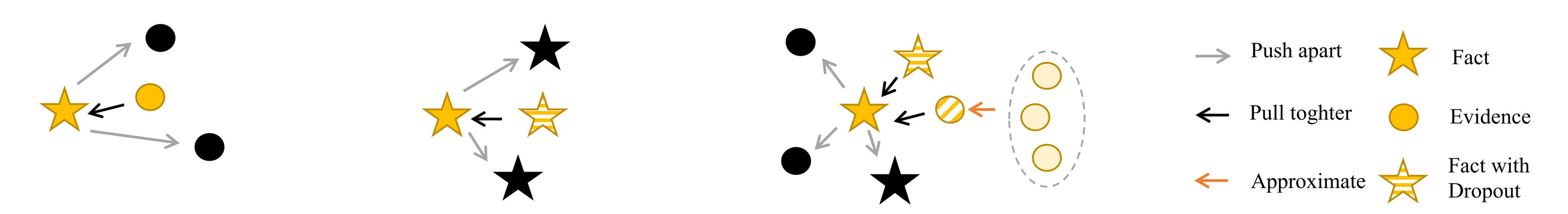
Due to the high cost of annotating fine-grained fact-evidence similarity, we propose to solve this task in an unsupervised way. Therefore, we constructed two types of dataset: 1) **Unsupervised Dataset.** a coarse parallel fact-evidence aligned corpus; 2) **Supervised Dataset.** a dataset with fine-grained relevance ranking annotations.

Task	#Query	#Can./que.	#Que-can. pair	#Char./que.	#Pos./que.	#Case	#Crime	Granularity
LeCaRD	107	100	10,700	444.58	10.33	10,700	20	Document-level
LERD-usp	308,749	35	11,079,998	63.22	–	35,423	255	Sentence-level
LERD-sup	4,336	54	234,693	67.63	3.15	919	91	Sentence-level

Methods

SWAP: Structure-aWare contrastive learning with Approximate aggregated Positive

We formulate the LER task as Dense Retrieval without supervision. Specifically, we proposed a positive sample constructing strategy using the case structure information, which is also used to do negative sampling.



(a) Supervised Constrastive Learning
(b) Self-supervised Constrastive Learning
(c) Constrastive Learning with Approximate Positive

Supervised Setting: given a fact f_i, its relevant evidence e_i^+ and a set of r negative $E_i^- = \{e_{i,j}^-\}^r$, the loss is $\mathcal{L}^{\mathrm{C}}$.

$$\mathcal{L}^{\mathrm{C}} = -log\frac{\exp^{sim(f_i,e_i^+)/\tau}}{\exp^{sim(f_i,e_i^+)/\tau}+\sum_{j=1}^{r}\exp^{sim(f_i,e_{i,j}^-)/\tau}}; \qquad \hat{a}_i^{k+} = \sum_{j=1}^{n}\frac{\mathrm{e}^{\hat{f}_i^k\cdot\hat{e}_j^k}}{\sum_{l=1}^{m}\mathrm{e}^{\hat{f}_i^k\cdot\hat{e}_l^k}}\cdot\hat{e}_j^k;$$

Unsupervised Setting: the positive e_i^+ for f_i is not given, but we notice the true positive e_i^{k+} for f_i^k is doomed to be within the evidence collection $E_k = \{e_j^k\}_{j=1}^n$ from the same case k by nature. Thus, we approximate a_i^{k+} via aggregating the representations of all e_j^k in $E_k = \{e_j^k\}_{j=1}^n$. We further mitigate the approximation error by denoising.

$$L_{a_i^k}^{\mathrm{DE}} = -\log\frac{w_i^k\cdot\mathrm{e}^{sim(f_i^k,a_i^{k+})/\tau}}{w_i^k\cdot\mathrm{e}^{sim(f_i^k,a_i^{k+})/\tau}+\sum_{a_j\in\mathcal{U}_{a_i^k}}w_j^k\cdot\mathrm{e}^{sim(f_i^k,a_j)/\tau}}; \quad w_i^k = \sqrt{\sum_{j=1}^{n}\left(\frac{\mathrm{e}^{f_i^k\cdot e_j^k}}{\sum_{l=1}^{n}\mathrm{e}^{f_i^k\cdot e_l^k}}\right)^2}$$

Main Results

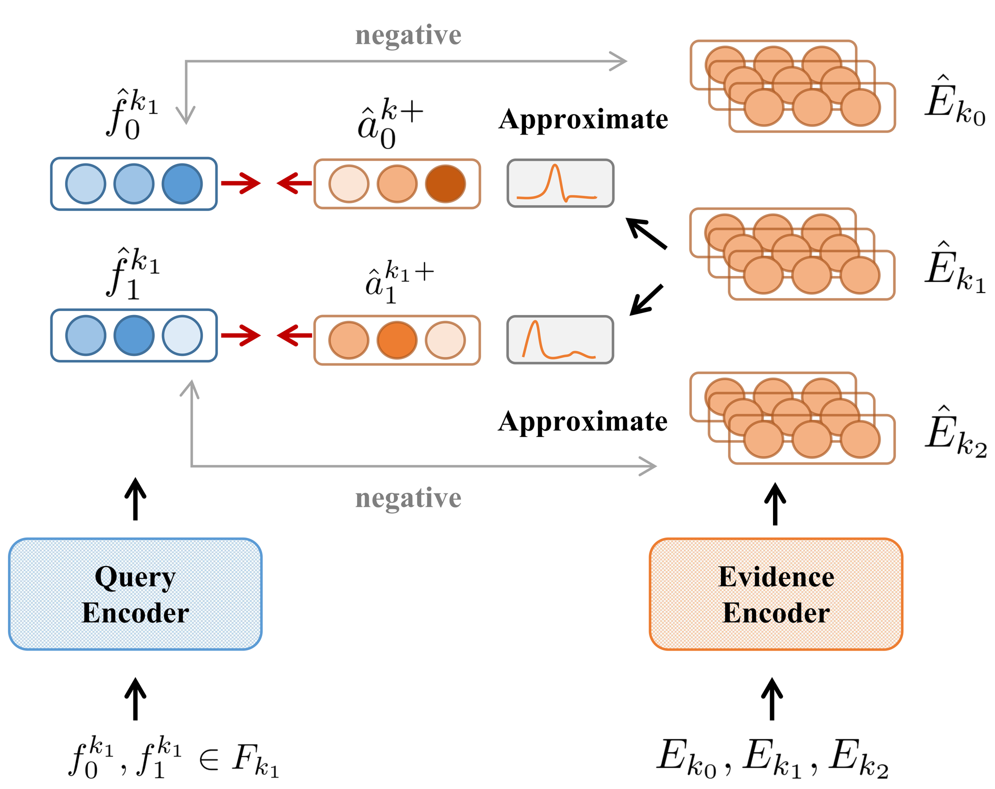
Category	Method	MAP	MRR	R@1	R@3	R@5	NDCG@1	NDCG@3	NDCG@5
Sparse Retrieval	BM25	39.03	45.83	29.03	38.20	48.29	31.10	35.29	39.75
	Legal-Event-IR	37.25	45.29	30.12	36.71	44.52	32.45	35.15	38.49
Text Representation	BERT	47.97	58.15	43.77	49.46	57.25	46.16	47.64	50.58
	RoBERTa	51.63	61.62	47.51	53.11	61.44	50.10	51.28	54.58
	LawFormer	52.25	62.52	48.78	54.05	61.73	51.40	52.42	55.23
	SBERT°	40.51	50.13	34.39	40.52	49.19	36.63	38.61	42.11
	SimCSE*	56.09	66.39	52.99	58.73	66.06	55.60	56.54	59.29
Dense Retrieval	Contriever°	45.44	56.11	41.37	46.56	55.18	43.64	44.65	48.08
	Contriever(MS)°	53.67	64.68	50.89	55.52	64.02	53.50	53.88	57.09
	Condensor*	54.65	64.82	51.01	57.14	64.40	53.57	55.08	57.80
	SWAP-BERT(ours)	59.65	69.58	56.68	62.09	69.83	59.44	60.43	63.33
	SWAP-RoBERTa(ours)	**61.45**	**71.25**	**58.92**	**64.11**	**71.97**	**61.62**	**62.34**	**65.27**
Supervised	DPR-RoBERTa	62.07	72.94	60.02	64.91	70.04	63.06	63.50	63.99
	DPR-SWAP-RoBERTa	**64.07**	**75.67**	**64.05**	**67.92**	**73.44**	**66.82**	**66.22**	**67.64**

Experiment Setups

- **Dataset Statistics**

Setting	Split	#Query	#Que-can pair.	#Case	#Crime
Usp	train	308,749	11,079,998	35,423	255
	valid	943	57,017	200	44
	test	3,393	177,676	719	84
Sup	train	2,940	154,257	619	78
	valid	453	23,419	100	34
	test	943	57,017	200	44

- **Model Architecture**



Performance Analysis

- **Effect of Each Component**

Method	MAP	MRR	R@5	NDCG@5
SWAP	61.45	71.25	71.97	65.27
SWAP$_{wo\text{-}DE}$	60.72	70.57	70.56	64.09
−AP	51.01	61.98	59.39	53.92
−DP	52.90	62.89	63.16	56.03
−DP-in-case	54.81	64.72	65.37	58.15
SWAP^{cls}	57.10	67.39	67.00	60.78
SWAP$^{cls}_{wo\text{-}DE}$	56.77	66.94	66.02	59.97
−AP	45.17	55.69	54.13	47.79
−DP	48.18	58.05	57.89	51.13
−DP-in-case	51.01	60.93	60.84	54.02

- **Effect of Backbone**

Backbone	setting	MAP	MRR	R@5	NDCG@5
Bert-tiny	vanilla	45.82	55.77	54.99	48.15
	SWAP$_{wo\text{-}DE}$	51.48	61.39	60.99	54.17
	SWAP	52.17	62.38	61.85	55.00
Bert	vanilla	47.97	58.15	57.25	50.58
	SWAP$_{wo\text{-}DE}$	57.42	67.93	67.67	61.14
	SWAP	59.65	69.58	69.83	63.33
Roberta	vanilla	51.63	61.62	61.44	54.58
	SWAP$_{wo\text{-}DE}$	60.72	70.57	70.56	64.09
	SWAP	61.45	71.25	71.97	65.27
Mengzi	vanilla	48.91	59.25	57.90	51.47
	SWAP$_{wo\text{-}DE}$	61.17	70.98	71.78	64.94
	SWAP	61.31	71.32	71.16	65.06
Ernie	vanilla	47.03	56.89	55.98	49.46
	SWAP$_{wo\text{-}DE}$	59.29	69.68	68.61	62.77
	SWAP	60.83	70.34	71.34	64.71

- **Effect of Data Scale & Type**

Train	MAP	MRR	R@5	NDCG@5
1K	59.17	69.18	69.21	62.48
3K	60.31	70.12	70.54	63.90
All	61.45	71.25	71.97	65.27
Drug	55.06	65.81	64.87	58.22
Steal	59.21	69.45	69.36	62.61
Bodily-harm	58.42	68.85	68.92	62.15