DocRED: A Large-Scale Document-Level Relation Extraction Dataset

소프트웨어학부 2017012333 이수아

CONTETNS

연구 주제

요약

실험

결론

연구주제

Relation Extraction(RE)

identify relational facts between entities from plain text

Kungliga Hovkapellet

[1] ***Kungliga Hovkapellet*** (The ***Royal Court Orchestra***) is a *Swedish* orchestra, originally part of the *Royal Court* in ***Sweden***'s capital *Stockholm*. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until *1727*, when *Sophia Schröder* and *Judith Fischer* were employed as vocalists; in the *1850s*, the harpist *Marie Pauline Åhman* became the first female instrumentalist. [4] From *1731*, public concerts were performed at ***Riddarhuset*** in *Stockholm*. [5] Since *1773*, when the ***Royal Swedish Opera*** was founded by *Gustav III* of ***Sweden***, the ***Kungliga Hovkapellet*** has been part of the opera's company.

Subject: ***Kungliga Hovkapellet; Royal Court Orchestra***
Object: ***Royal Swedish Opera***
Relation: `part_of` **Supporting Evidence:** 5

Subject: ***Riddarhuset***
Object: ***Sweden***
Relation: `country` **Supporting Evidence:** 1, 4

Sentense-level RE

Document-level RE

연구주제

Document-level RE

large-scale annotated dataset이 필요!

Problems		*Solutions*
- small # of manually-annotated relations and entities	→	large-scale
- exhibit noisy annotations from distant supervision	→	manually-annotated
- serve specific domains or approaches	→	general-purpose

연구주제

DocRED: *Document-Level Relation Extraction Dataset*

large-scale human-annotated document-level RE dataset

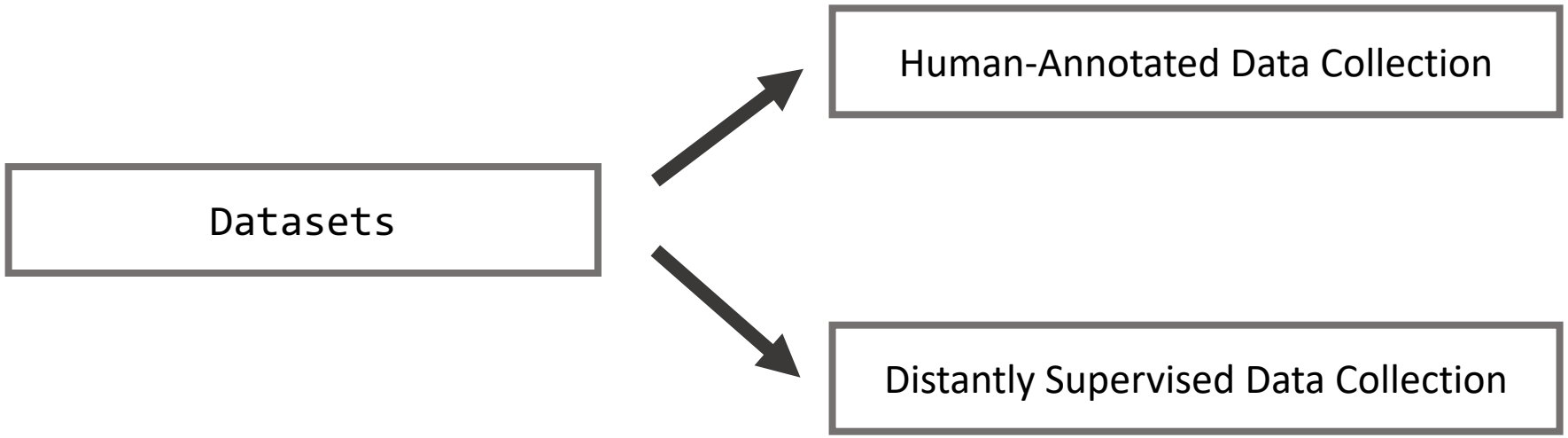
- 132,375 entities and 56,354 relational facts annotated on 5,053 Wikipedia documents, making it the largest human-annotated document-level RE dataset.

- As at least 40.7% of the relational facts in DocRED can only be extracted from multiple sentences, DocRED requires reading multiple sentences in a document to recognize entities and inferring their relations by synthesizing all information of the document.

- We also provide large-scale distantly supervised data to support weakly supervised RE research.

요약

Data Collection



요약

Human-Annotated Data Collection

Stage 1: Distantly Supervised Annotation Generation

➡ spaCy - named entity recognition – linked to Wikidata items – relations labeling

➡ 107,050 documents / 5,053 documents & 96 relations for human annotation

Stage 2: Named Entity and Coreference Annotation

Stage 3: Entity Linking

Stage 4: Relation and Supporting Evidence Collection

➡ recommendations from RE models & distant supervision based on entity linking

요약

Distantly Supervised Data Collection

101,873 documents

Named entity mentions are reidentified BERT

link each named entity mention to one Wikidata item by a heuristic-based method

merge the named entity mentions with identical KB IDs

relations between each merged entity pair are labeled via distant supervision

요약

Data Analysis

Dataset	# Doc.	# Word	# Sent.	# Ent.	# Rel.	# Inst.	# Fact
SemEval-2010 Task 8	-	205k	10,717	21,434	9	8,853	8,383
ACE 2003-2004	-	297k	12,783	46,108	24	16,771	16,536
TACRED	-	1,823k	53,791	152,527	41	21,773	5,976
FewRel	-	1,397k	56,109	72,124	100	70,000	55,803
BC5CDR	1,500	282k	11,089	29,271	1	3,116	2,434
DocRED (Human-annotated)	5,053	1,002k	40,276	132,375	96	63,427	56,354
DocRED (Distantly Supervised)	101,873	21,368k	828,115	2,558,350	96	1,508,320	881,298

요약

Data Analysis

Reasoning Types	%	Examples
Pattern recognition	38.9	[1] ***Me Musical Nephews*** is a **1942** one-reel animated cartoon directed by Seymour Kneitel and animated by Tom Johnson and George Germanetti. [2] Jack Mercer and Jack Ward wrote the script. ... **Relation:** `publication_date` **Supporting Evidence:** 1
Logical reasoning	26.6	[1] "Nisei" is the ninth episode of the third season of the American science fiction television series The X-Files. ... [3] It was directed by David Nutter, and written by **Chris Carter**, Frank Spotnitz and Howard Gordon. ... [8] The show centers on FBI special agents ***Fox Mulder*** (David Duchovny) and Dana Scully (Gillian Anderson) who work on cases linked to the paranormal, called X-Files. ... **Relation:** `creator` **Supporting Evidence:** 1, 3, 8
Coreference reasoning	17.6	[1] ***Dwight Tillery*** is an American politician of the Democratic Party who is active in local politics of Cincinnati, Ohio. ... [3] He also holds a law degree from the **University of Michigan Law School**. [4] ***Tillery*** served as mayor of Cincinnati from 1991 to 1993. **Relation:** `educated_at` **Supporting Evidence:** 1, 3
Common-sense reasoning	16.6	[1] ***William Busac*** (1020-1076), son of William I, Count of Eu, and his wife Lesceline. ... [4] ***William*** appealed to King Henry I of France, who gave him in marriage **Adelaide**, the heiress of the county of Soissons. [5] **Adelaide** was daughter of Renaud I, Count of Soissons, and Grand Master of the Hotel de France. ... [7] ***William*** and **Adelaide** had four children: ... **Relation:** `spouse` **Supporting Evidence:** 4, 7

Table 2: Types of reasoning required for document-level RE on DocRED. The rest 0.3% requires other types of reasoning, such as temporal reasoning. The ***head***, **tail** and `relation` are colored accordingly.

reasoning is essential
for document-level RE.

modeling interactions
between multiple entities

요약

Benchmark Settings

Setting		# Doc.	# Rel.	# Inst.	# Fact
Train	W	101,873	96	1,508,320	881,298
	S	3,053	96	38,269	34,715
Dev	S,W	1,000	96	12,332	11,790
Test	S,W	1,000	96	12,842	12,101

Table 3: Statistics of data used for the two benchmark settings (Sec. 4): supervised setting (S) and weakly supervised setting (W).

요약

Benchmark Settings

Supervised Setting

- only human-annotated data is used
- rich reasoning skills

Weakly Supervised Setting

- Training set : distantly supervised data
- wrong labeling problem accompanied with distantly supervised data

결론

Experiment

Models - CNN, LSTM, Bidirectional LSTM, Context-Aware based model

$\mathcal{D} = \{w_i\}_{i=1}^{n}$ D: Document

$\{\boldsymbol{h}_i\}_{i=1}^{n}$ h_i: hidden state vector sequence

$\boldsymbol{m}_k = \frac{1}{t-s+1}\sum_{j=s}^{t}\boldsymbol{h}_j$ m_k: named entity mention

$\boldsymbol{e}_i = \frac{1}{K}\sum_k \boldsymbol{m}_k$ e_i : entity

$$\hat{\boldsymbol{e}}_i = [\boldsymbol{e}_i; \boldsymbol{E}(d_{ij})],\ \hat{\boldsymbol{e}}_j = [\boldsymbol{e}_j; \boldsymbol{E}(d_{ji})]$$

$$P(r|e_i, e_j) = \text{sigmoid}(\hat{\boldsymbol{e}}_i^{\mathrm{T}}\boldsymbol{W}_r\hat{\boldsymbol{e}}_j + b_r)$$

d_{ij}, d_{ji}: 두 entity에 대한 relative distances
E : embedding matrix ; r : relation type
W_r, b_r : relation type에 따른 trainable parameters

결론

Experiment

Models – CNN, LSTM, Bidirectional LSTM, Context-Aware based model

Model	Dev				Test			
	Ign F1	Ign AUC	F1	AUC	Ign F1	Ign AUC	F1	AUC
Supervised Setting								
CNN	41.58	36.85	43.45	39.39	40.33	36.24	42.26	38.91
LSTM	48.44	46.62	50.68	49.48	47.71	46.27	50.07	49.25
BiLSTM	48.87	**47.61**	50.94	**50.26**	**48.78**	**47.61**	**51.06**	**50.43**
Context-Aware	**48.94**	47.22	**51.09**	50.17	48.40	46.54	50.70	49.64
Weakly Supervised Setting								
CNN	33.24	23.17	42.76	37.99	32.33	21.83	42.00	36.84
LSTM	39.37	22.39	49.92	42.79	38.27	21.74	48.88	41.35
BiLSTM	**41.44**	**23.21**	**51.72**	**44.44**	39.15	**22.14**	49.80	**42.87**
Context-Aware	40.47	22.56	51.39	43.00	**39.16**	21.58	**50.12**	41.51

Table 4: Performance of different RE models on DocRED (%).

결론

Experiment

Method	RE			RE+Sup		
	P	R	F1	P	R	F1
Model	55.6	52.6	54.1	46.4	43.1	44.7
Human	**89.7**	**86.3**	**88.0**	**71.2**	**75.8**	**73.4**

Table 5: Human performance (%).

Method	Dev	Test
Heuristic predictor	36.21	36.76
Neural predictor	**44.07**	**43.83**

Table 7: Performance of joint relation and supporting evidence prediction in F1 measurement (%).

결론

Conclusion

Exploring models explicitly considering reasoning

Designing more expressive model architectures for collecting and synthesizing inter-sentence information

Leveraging distantly supervised data to improve the performance of document-level RE

document-level RE > sentence-level RE

끝