Exploring Partial Knowledge Base Inference in Biomedical Entity Linking

Hongyi Yuan[1*]; Keming Lu[2*]; Zheng Yuan[3#]
[1]Tsinghua University, [2]University of Southern California, [3]Alibaba Group
*Contributed Equally, #Corresponding Author

Abstract

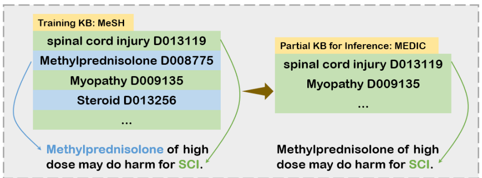
Entity Linking models are trained on corpora labeled by a predefined KB. However, it is a common scenario that only entities within a subset of the KB are precious to stakeholders. We explore the practical scenario named partial knowledge base inference: training an EL model with one KB and inferring on the part of it without further training. In this work, we give a detailed definition and evaluation procedures for the scenario.

By deliberately constructed benchmarks, we witness degradation in performance when inference on partial KB which reveals that existing EL paradigms can not handle unlinkable mentions (NIL) correctly. We explore two simple-and-effective redemption methods to combat the NIL issue with little computational overhead.

Codes are released at https://github.com/Yuanhy1997/PartialKB-EL.

Introduction

Task Definition: an example visual illustration.

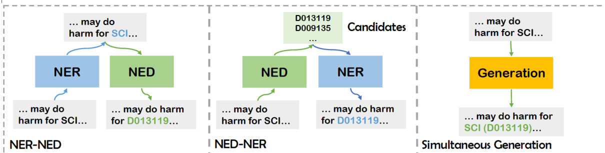


Partial KB inference from *training KB* **MeSH** to a *partial KB* **MEDIC** (Note that MEDIC is a sub-KB of MeSH). ***Methylprednisolone*** should not be linked since it is not in MEDIC. This represents a practical scenario where different stakeholders may only concern entities from a special KB.

For EL, it is to develop a mapping from a text to a set of mention-concept pairs: $f: s \to \mathcal{P}_{\mathcal{E}}$, where $\mathcal{P}_{\mathcal{E}} = \{(i,j,e) | 0 \le i \le j, e \in \mathcal{E}\}$. $\mathcal{E}$ is a KB.

For Partial KB inference, we assume models trained on annotations $\mathcal{P}_{\mathcal{E}_1}$ while inference only concern annotations $\mathcal{P}_{\mathcal{E}_2}$. We assume $\mathcal{E}_1 \supset \mathcal{E}_2$.

Three existing EL paradigms:



NER-NED is a pipeline paradigms where first an NER methods detect all the mentions in texts and then an ED method links each mention to a concept.
NED-NER firstly retrieves all the possible concepts in texts and then ground each mention by a possible concept. E.g. EntQA
Simultaneous Generation is a generative method where generate mentions and concept names autoregressively. E.g. GENRE

Probing Methods and Materials

We use two popular biomedical EL datasets: MedMentions and BC5CDR.

MedMentions: all annotations corresponding to UMLS as training KB, we use SNOMED and two semantic type T038 and T058 and their complements as partial KBs.

BC5CDR: all annotations corresponding to MeSH as training KB, we use MEDIC and its complements as a partial KB.

For NER-NED we use a combination of KeBioLM and CODER; for NED-NER we use EntQA and for simultaneous generation we use GENRE.

Probing Results

By assessing overall linking performance and decomposed performance of NER and NED, which we come with several observation:

1. NER-NED and simultaneous generation methods suffer from a large degradation, while NED-NER method is robust to the scenario.
2. The performance degradation mainly caused by a large degradation in NER, where a large percent of NILs are recalled causing a sharp drop in Precision.

Overall Linking Performance:

	Target KB.		EntQA			GENRE			KeBioLM+CODER		
	Train KB	Eval KB	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
MedMentions	UMLS	UMLS	45.99	23.68	31.27	42.44	43.69	43.05	33.58	34.94	34.25
		SNOMED	46.04	27.01	34.05	34.40	49.40	40.56	28.19	48.28	35.59
		$SNOMED^C$	36.75	23.12	28.38	19.82	39.28	26.35	14.18	37.54	20.59
		T038	41.52	31.56	35.86	17.26	49.53	25.60	9.78	50.28	16.37
		$T038^C$	43.43	23.24	30.28	34.97	42.45	38.35	26.52	34.59	30.02
		T058	30.01	25.56	27.61	7.69	36.06	12.68	4.76	41.51	8.54
		$T058^C$	46.02	24.34	31.84	40.45	44.76	42.50	31.95	37.74	34.61
	Avg. Drop		5.36	-2.13	-0.7	16.68	0.11	12.04	14.35	-6.71	9.96
BC5.	MeSH	MeSH	83.59	66.48	74.06	70.92	68.71	69.80	72.21	74.84	73.5
		MEDIC	81.92	70.45	75.75	31.53	68.19	43.12	29.24	68.38	40.96
		$MEDIC^C$	87.10	66.92	75.69	37.55	65.33	47.69	42.57	80.67	55.73
	Avg. Drop		-0.92	-2.21	-1.66	36.38	1.95	24.40	36.31	0.32	25.16

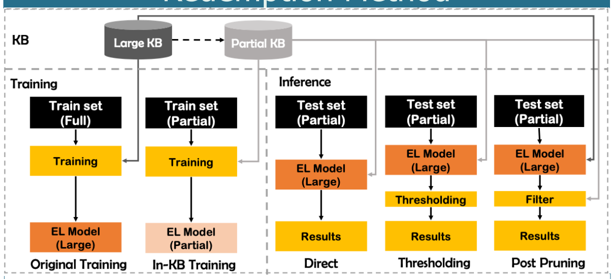
NER performance:

	Target KB.		EntQA			GENRE			KeBioLM+CODER		
	Train KB	Eval KB	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Medmentions	UMLS	UMLS	82.72	51.81	63.72	64.27	66.17	65.21	69.08	71.88	70.45
		SNOMED	82.09	51.83	63.54	45.78	65.74	53.97	43.22	74.04	54.58
		$SNOMED^C$	80.57	53.82	64.53	30.34	60.13	40.33	26.94	71.32	39.11
		T038	82.43	52.66	64.27	22.34	64.10	33.13	14.85	76.36	24.86
		$T038^C$	82.08	51.83	63.54	53.44	64.86	58.60	55.26	72.07	62.56
		T058	78.92	56.54	65.88	11.37	53.31	18.75	7.76	67.68	13.93
		$T058^C$	82.91	50.53	62.78	59.90	66.30	62.94	62.35	73.65	67.53
	Avg. Drop		1.22	-1.06	-0.37	27.08	3.76	20.59	34.02	-0.64	26.68
BC5.	MeSH	MeSH	94.67	82.56	88.20	87.59	84.86	86.20	86.47	91.05	88.70
		MEDIC	92.31	84.04	87.99	37.85	81.84	51.76	36.46	86.46	51.29
		$MEDIC^C$	96.37	82.93	89.14	49.07	85.38	62.32	50.13	94.99	65.63
	Avg. Drop		0.33	-0.93	-0.37	44.13	1.25	29.16	43.18	0.33	30.24

NED performance:

	Target KB.		EntQA		GENRE	Ke.+CO.
	Train KB	Eval KB	R@100	Acc.	Acc.	Acc.
MedMentions	UMLS	UMLS	57.26	75.38	66.03	48.61
		SNOMED	65.86	74.81	75.14	65.22
		$SNOMED^C$	61.72	68.67	65.33	52.64
		T038	75.10	65.34	77.26	65.86
		$T038^C$	58.54	66.89	65.44	47.99
		T058	74.28	57.92	67.63	61.34
		$T058^C$	58.76	68.52	67.53	51.24
	Avg. Drop		-8.45	8.35	-3.69	-8.77
BC5.	MeSH	MeSH	80.34	92.72	80.97	83.51
		MEDIC	88.72	90.95	83.30	80.20
		$MEDIC^C$	77.73	93.23	76.52	84.92
	Avg. Drop		-2.89	0.63	1.06	0.95

Redemption Method



We explore two redemption methods: Thresholding and Posting Pruning.
Thresholding screens NILs by models' confidence score;
Posting Pruning screens NILs by omitting those results not in partial KBs.

		MEDIC				$MEDIC^C$			
		EL-P/R	EL-F1	NER-F1	NED-Acc	EL-P/R	EL-F1	NER-F1	NED-Acc
EntQA	In-KB Train	81.27/71.34	**75.98**	**88.16**	**92.14**	86.87/69.30	**77.10**	**90.08**	**94.44**
	Partial KB Inference	81.92/70.45	75.75	87.99	90.95	87.10/66.92	75.69	89.14	93.23
	w/ Post-pruning	62.97/64.99	63.96	84.10	80.76	80.02/63.11	70.57	86.42	77.84
GENRE	In-KB Train	65.65/68.38	66.99	78.56	85.26	69.96/62.02	65.75	85.52	76.89
	Partial KB Inference	31.53/68.19	43.12	51.76	83.30	37.55/65.33	47.69	62.32	76.52
	w/ Thresholding	76.32/59.25	66.71	72.43	**92.11**	69.05/56.99	62.45	74.86	**83.41**
	w/ Post-pruning	69.31/68.59	**68.95**	**79.92**	86.27	69.46/66.29	**67.83**	**86.47**	78.45
Ke.+CO.	In-KB Train	63.98/68.47	66.15	**82.94**	80.48	77.52/80.65	79.05	**92.82**	85.18
	Partial KB Inference	29.24/68.38	40.96	51.29	80.20	42.57/80.67	55.73	65.63	84.92
	w/ Thresholding	79.20/65.08	**71.45**	78.46	**91.07**	86.32/77.04	**81.41**	83.35	**97.68**
	w/ Post-pruning	69.03/65.27	67.10	78.48	85.49	69.17/80.67	74.48	87.27	85.34

Conclusions

In this research piece, a practical scenario called partial KB inference is explored and we empirically show that existing EL paradigms degrade under trivial transferring. We also propose two different redemption methods named thresholding and post-pruning where both methods bring improvement to the scenario.
Our findings illustrate the importance of partial KB inference in EL.

Contact

Hongyi Yuan, Keming Lu and Zheng Yuan
Tsinghua University, University of Southern California and Alibaba Group

yuanhy20@mails.tsinghua.edu.cn
keminglu@usc.edu
yuanzheng.yuanzhen@alibaba-inc.com

References

De Cao N, Izacard G, Riedel S, et al. Autoregressive entity retrieval[J]. arXiv preprint arXiv:2010.00904, 2020.
Zhang W, Hua W, Stratos K. EntQA: Entity linking as question answering[J]. arXiv preprint arXiv:2110.02369, 2021.
Yuan Z, Zhao Z, Sun H, et al. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization[J]. Journal of biomedical informatics, 2022, 126: 103983.
Yuan Z, Liu Y, Tan C, et al. Improving biomedical pretrained language models with knowledge[J]. arXiv preprint arXiv:2104.10344, 2021.
Mohan S, Li D. Medmentions: A large biomedical corpus annotated with umls concepts[J]. arXiv preprint arXiv:1902.09476, 2019.
Li J, Sun Y, Johnson R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction[J]. Database, 2016, 2016.
Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology[J]. Nucleic acids research, 2004, 32(suppl_1): D267-D270.
Stearns M Q, Price C, Spackman K A, et al. SNOMED clinical terms: overview of the development process and project status[C]//Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001: 662.
Wu W H, Bui A A T, Batalin M A, et al. MEDIC: Medical embedded device for individualized care[J]. Artificial intelligence in medicine, 2008, 42(2): 137-152.