Text Augmented Open Knowledge Graph Completion via Pre-Trained Language Models

Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun and Jiawei Han

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Code @

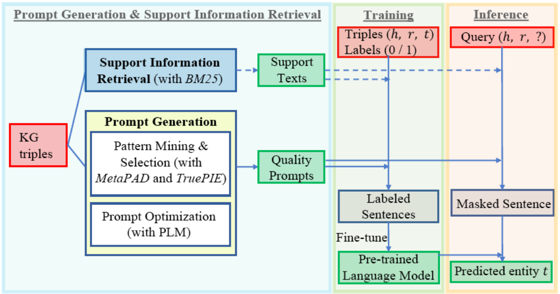
Introduction

- Knowledge Graphs (KGs) typically utilize related source corpora for the extraction of KG triples.
- Pre-trained Language Models (PLMs) can function effectively as knowledge bases.
- Existing research has leveraged PLMs for KG completion, primarily employing manually designed prompts - a process that can be resource-intensive in real-world situations.
- We introduce **TagReal**. Key features:

(1) Automates the process of identifying high-quality, inherent patterns within the corpus.
(2) Utilizes these discovered patterns as prompts for knowledge probing.
(3) Provides a more efficient and cost-effective solution for knowledge probing and KG completion.

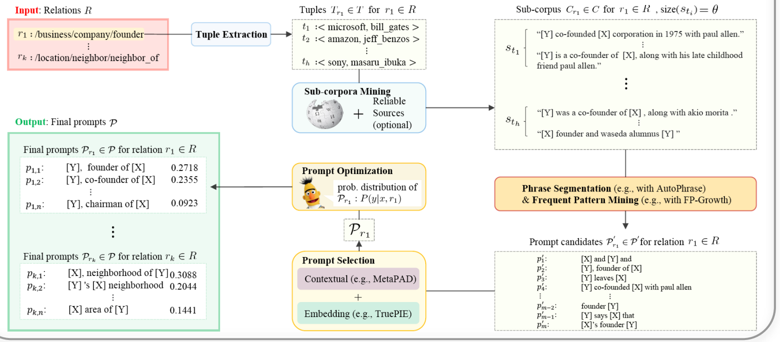
TagReal: Framework

- Two core module: prompt generation & support information retrieval



TagReal: Prompt Generation

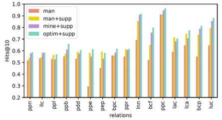
- Apply text/pattern mining methods for prompt mining
- An end-to-end solution to mine prompt from large corpus

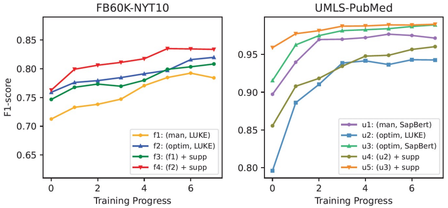


Results

	Model	20% Hits@5	20% Hits@10	20% MRR	50% Hits@5	50% Hits@10	50% MRR	100% Hits@5	100% Hits@10	100% MRR
KGE-based	TransE (Bordes et al., 2013)	29.13	32.67	15.80	41.54	45.74	25.82	42.53	46.77	29.86
	DisMult (Yang et al., 2014)	3.44	4.31	2.64	15.98	18.85	13.14	37.94	41.62	30.56
	ComplEx (Trouillon et al., 2016a)	4.32	5.48	3.16	15.00	17.73	12.21	35.42	38.85	28.59
	ConvE (Dettmers et al., 2018)	29.49	33.30	24.31	40.10	44.03	32.97	50.18	54.06	40.39
	TuckER (Balažević et al., 2019)	29.50	32.48	24.44	41.73	45.58	33.84	51.09	54.80	40.47
	RotatE (Sun et al., 2019)	15.91	18.32	12.65	35.48	39.42	28.92	**51.73**	55.27	**42.64**
Text&KGE-based	RC-Net (Xu et al., 2014)	13.48	15.37	13.26	14.87	16.54	14.63	14.69	16.34	14.41
	TransE+Line (Fu et al., 2019)	12.17	15.16	4.88	21.70	25.75	8.81	26.76	31.65	10.97
	JointNRE (Han et al., 2018)	16.93	20.74	11.39	26.96	31.54	21.24	42.02	47.33	32.68
RL-based	MINERVA (Das et al., 2017)	11.64	14.16	8.93	25.16	31.54	22.24	43.80	44.70	34.62
	CPL (Fu et al., 2019)	15.19	18.00	10.87	26.81	31.70	23.80	43.25	49.50	33.52
PLM-based	PKGC (Lv et al., 2022)	35.77	43.82	28.62	41.93	46.70	31.81	41.98	52.56	32.11
	TagReal (our method)	**45.59**	**51.34**	**35.41**	**48.98**	**55.64**	**38.03**	50.85	**60.64**	38.86

Condition	FB60K-NYT10 20%	FB60K-NYT10 50%	FB60K-NYT10 100%	UMLS-PubMed 20%	UMLS-PubMed 40%	UMLS-PubMed 70%	UMLS-PubMed 100%
man	(35.77, 43.82)	(41.93, 46.70)	(41.98, 52.56)	(31.08, 43.49)	(41.34, 52.44)	(47.39, 56.52)	(55.05, 59.43)
man+supp	(43.23, 47.74)	(47.10, 52.02)	(48.66, 57.46)	(32.95, 44.42)	(44.37, 54.96)	(51.98, 59.09)	(59.99, 61.23)
mine+supp	(44.54, 49.53)	(47.43, 53.87)	(49.03, 58.82)	(35.56, 45.33)	(45.35, 55.44)	(53.12, 59.65)	(60.27, 61.70)
optim+supp	(45.59, 51.34)	(48.98, 55.64)	(50.85, 60.64)	(35.83, 46.45)	(46.26, 55.99)	(53.46, 60.40)	(60.68, 62.88)





- TagReal significantly outperforms baselines especially with limited training data.
- Both prompt generation and support information retrieval have significant effects on boosting the KGC performance.
- Choice of PLM is important, especially for domain-specific KG datasets.

Findings & Future Directions

Findings:
- Inherent patterns in large corpora can serve as prompts for knowledge extraction from pre-trained language models.
- Text mining methods could provide a new avenue to analyze the workings of pre-trained language models.

Future Directions:
- Examine advanced text mining techniques for deeper analysis of language models.
- Explore potential cross-disciplinary collaborations between text mining and language model fields.