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

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

Department of Computer Science, University of Illinois at Urbana-Champaign

Presenter: Pengcheng (Patrick) Jiang
Email: pj20@Illinois.edu

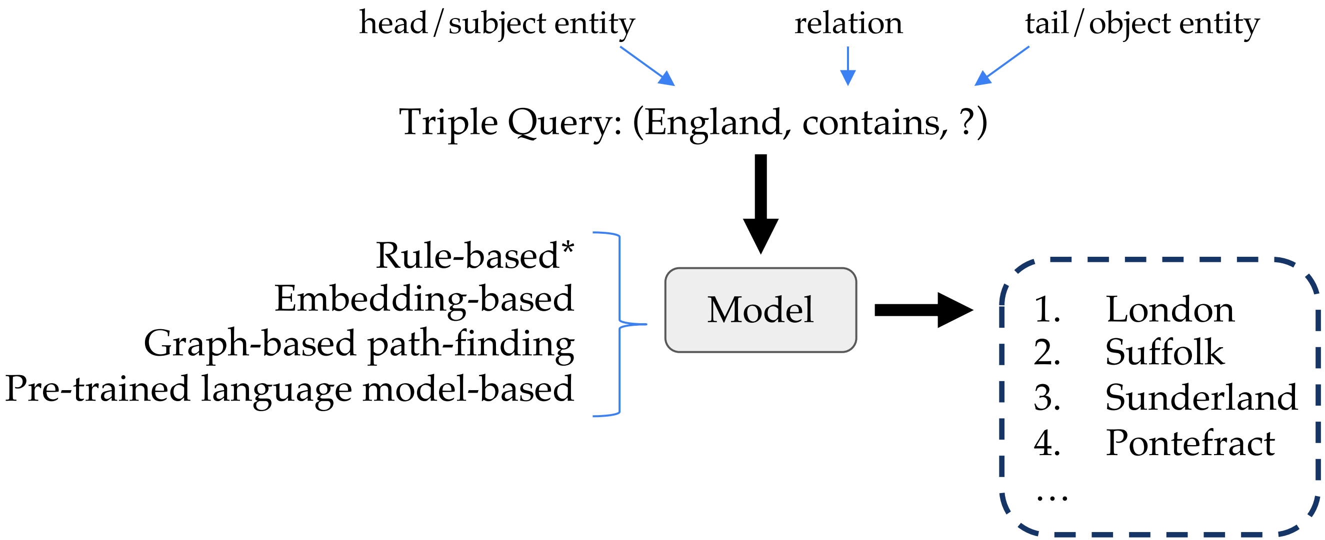
Overview

- Background & Motivation
- Methodology
- Experiments
- Conclusion & Thoughts

Background & Motivation

Background & Motivation

Task: **Knowledge Graph Completion (KGC)**



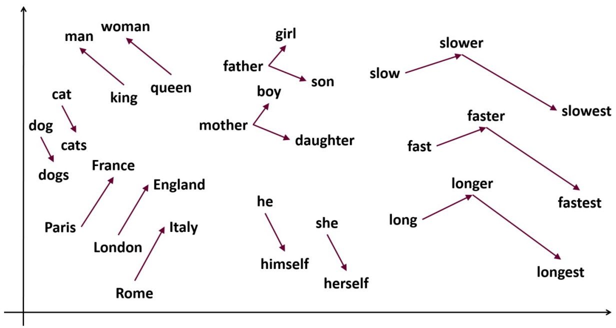
* We do not include rule-based methods in the discussion as their performance is no longer comparable to other methods.

Background & Motivation

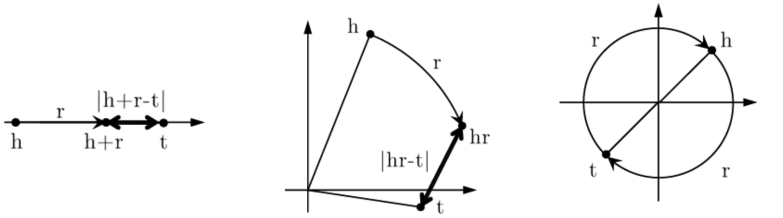
Early Stage: Knowledge Graph Embedding (KGE) models

(1) Translation-based models: TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), …
(2) Tensor-factorization based models: TuckER (BalaževiÅLc et al., 2019), HolE (Nickel et al., 2016), …
(3) Non-linear models: ConvE (Dettmers et al., 2018), ConvKB (Nguyen et al., 2017), …
(4) KGE with additional information: DKRL (Xie et al., 2016), KR-EAR (Lin et al., 2016), …

Mapping entities & relations into vector space



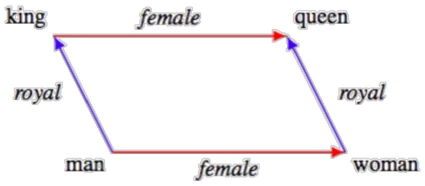
Define *score functions* as loss functions



(a) TransE models **r** as translation in real line.
(b) RotatE models **r** as rotation in complex plane.
(c) RotatE: an example of modeling symmetric relations **r** with $r_i = -1$

Figure 1: Illustrations of TransE and RotatE with only 1 dimension of embedding.

A well-known *example* (based on TransE):



(After training)
king + female ≈ queen
man + royal ≈ king
…

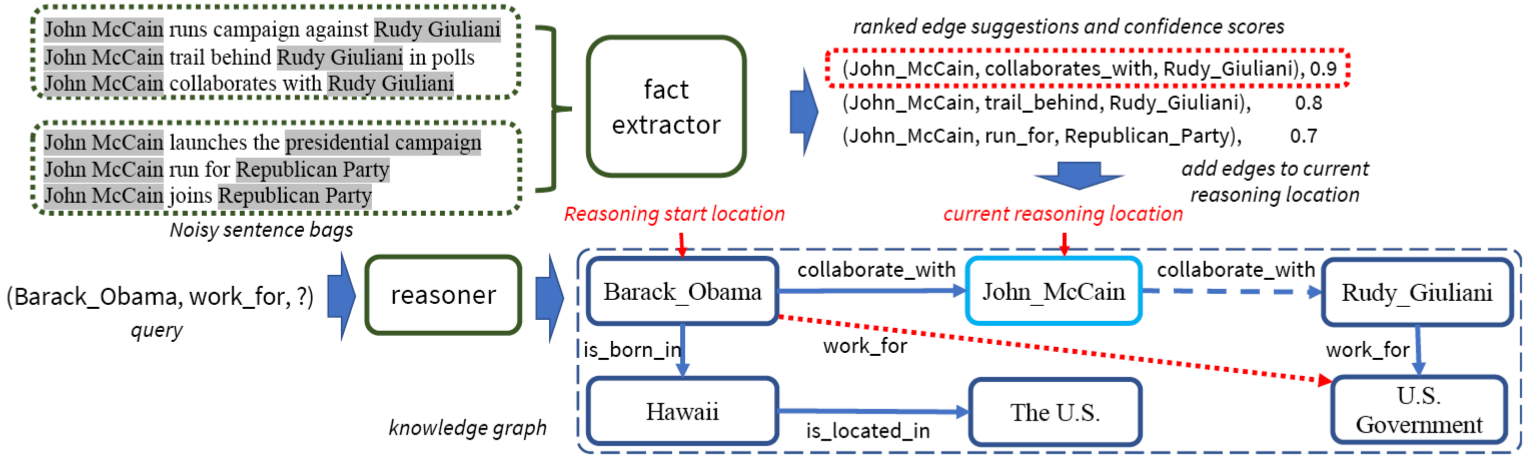
Limitations

1. **Need a huge amount of data for training**
2. **Not use the rich text corpus behind the KG**

Background & Motivation

Graph-based path-finding method

CPL* framework: (Complete the knowledge graph by finding an evidential path)



<u>Limitations</u>

Extracted set of facts is **noisy** and **constricted**
→ insufficient information to efficiently update the KG

* Fu, Cong, et al. "Collaborative Policy Learning for Open Knowledge Graph Reasoning." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Background & Motivation

Pre-trained Language Model-based Methods

Why PLM helps?

- researchers realize that pre-trained language models (PLM) can be knowledge bases

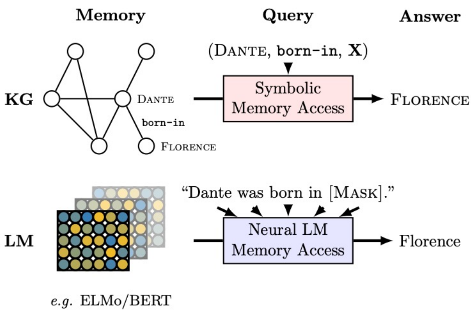


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], **Florence [-1.8]**, Naples [-1.9], Milan [-2.4], Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]**, London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], **dog [-2.4]**, cattle [-4.3], sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]**, French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], **midfielder [-2.4]**, forward [-2.4], midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], **Hamburg [-7.5]**, Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]**, Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
	P54	Dani Alves plays with ____.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], **sodium [-2.5]**, zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], **Uganda [-3.2]**, Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]**, Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]**, Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]**, Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
	P39	Pope Clement VII has the position of ____ .	pope	cardinal [-2.4], Pope [-2.5], **pope [-2.6]**, President [-3.1], Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], **Capitol [-3.2]**, Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]**, Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1], **ITV [-3.3]**, Global [-3.4], Frontier [-4.1], Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]**, Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], **Philippines [-3.6]**, February [-3.7], December [-3.8], Argentina [-4.0]

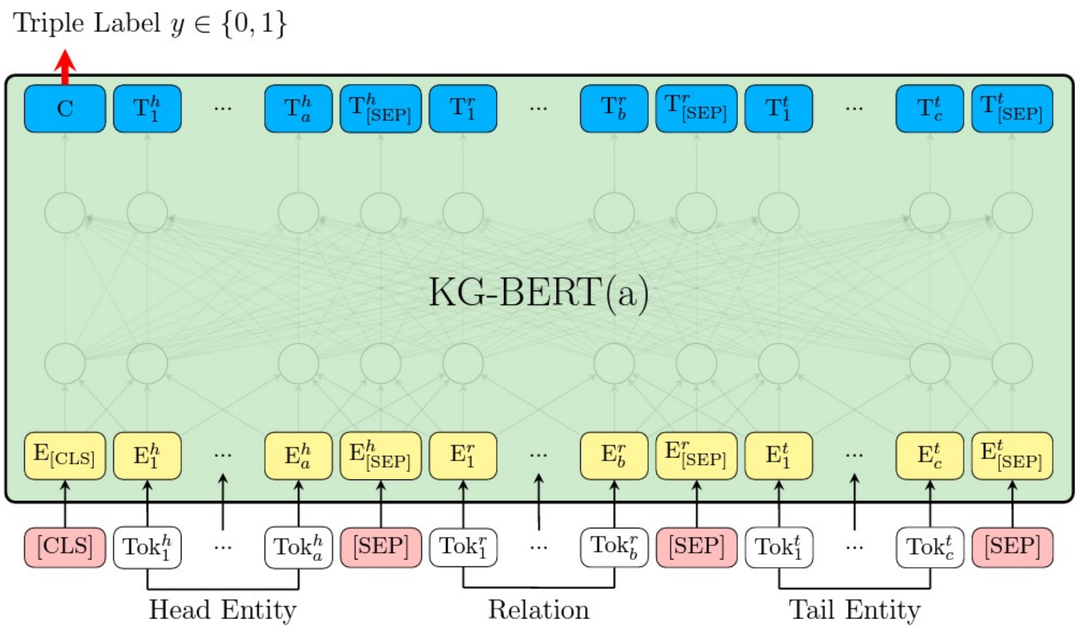
Knowledge Probing with BERT-Large

* Petroni, Fabio, et al. "Language Models as Knowledge Bases?." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Background & Motivation

Pre-trained Language Model-based Methods

KG-BERT* approach (Finetune a PLM with sliced triples)



Worse performance than KGE methods

Limitations

PLM is trained with a self-defined data type (sliced triple) s.t. the implicit knowledge in it could not be well exploited.

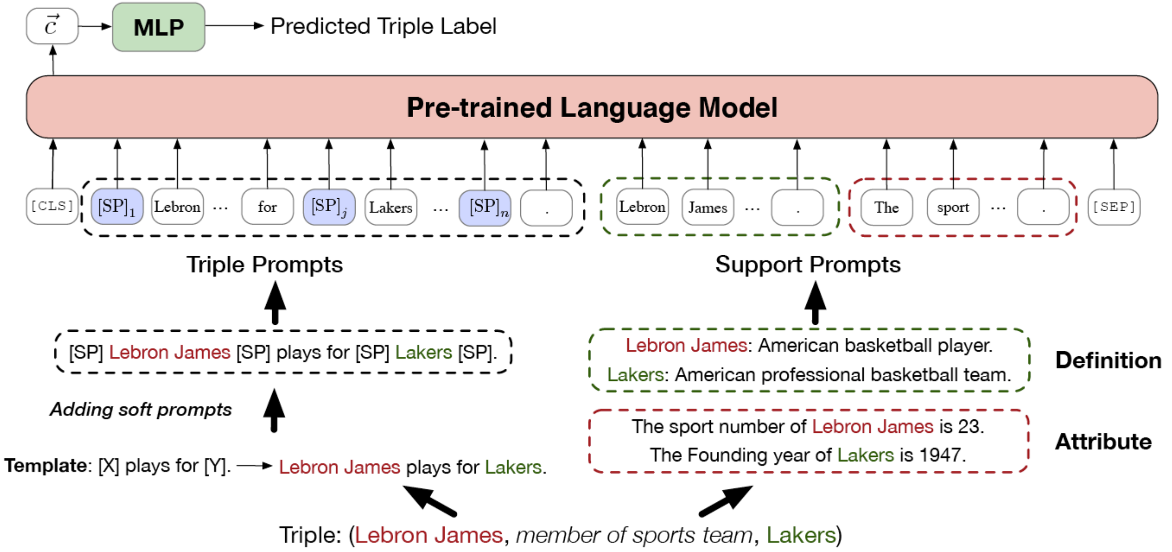
* Yao, Liang, Chengsheng Mao, and Yuan Luo. "KG-BERT: BERT for knowledge graph completion." *arXiv preprint arXiv:1909.03193* (2019).

Background & Motivation

Pre-trained Language Model-based Methods

PKGC* approach

- Transform KG triples to PLM-understandable text through prompt
- Append support information to fine-tune the PLM



Limitations

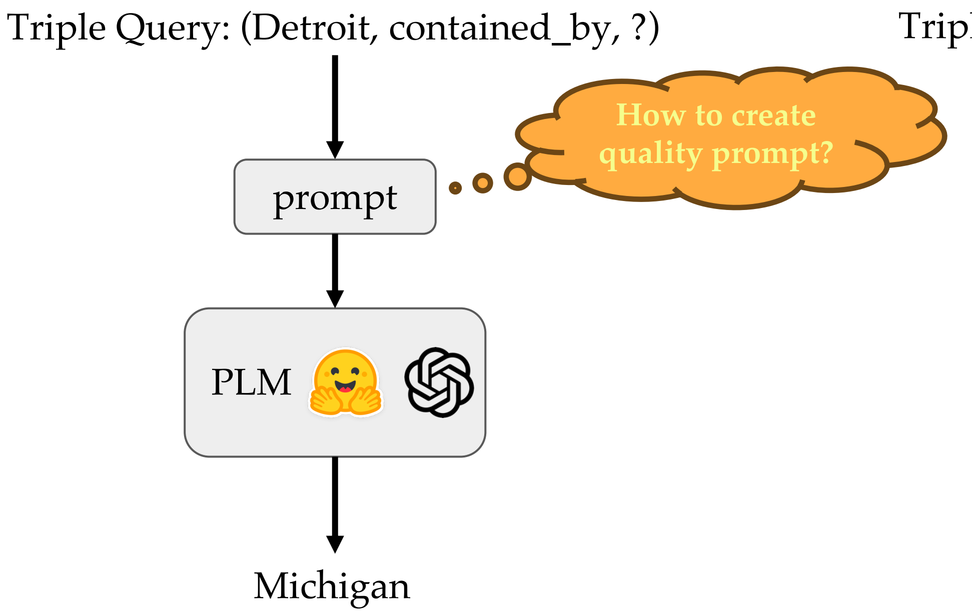
1. Human-powered prompt design is very costly.
2. The designed prompts may lead to suboptimal performance.
3. Support information is not pre-defined in most datasets

* Lv, Xin, et al. "Do pre-trained models benefit knowledge graph completion? a reliable evaluation and a reasonable approach." Association for Computational Linguistics (ACL), 2022.

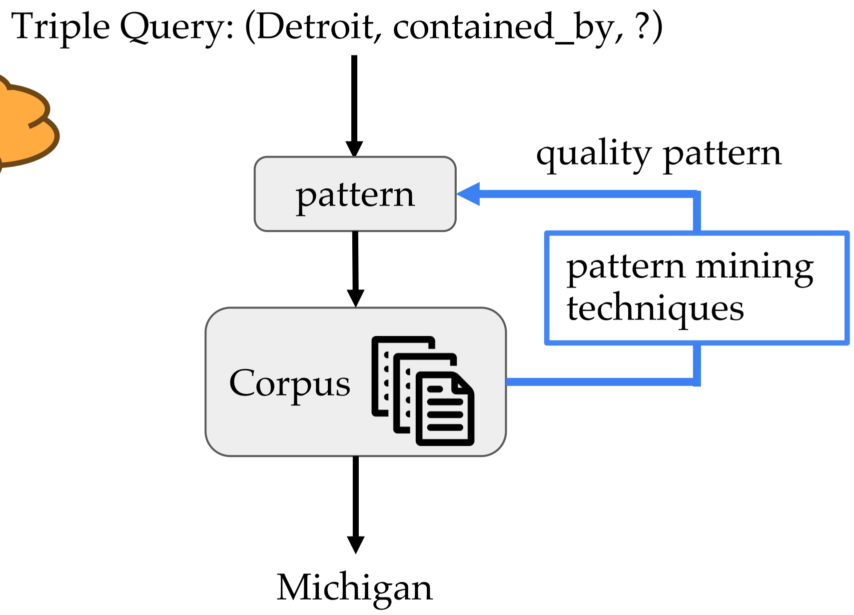
Background & Motivation

How can we generate quality prompt automatically?

<u>PLM-based KGC task</u>



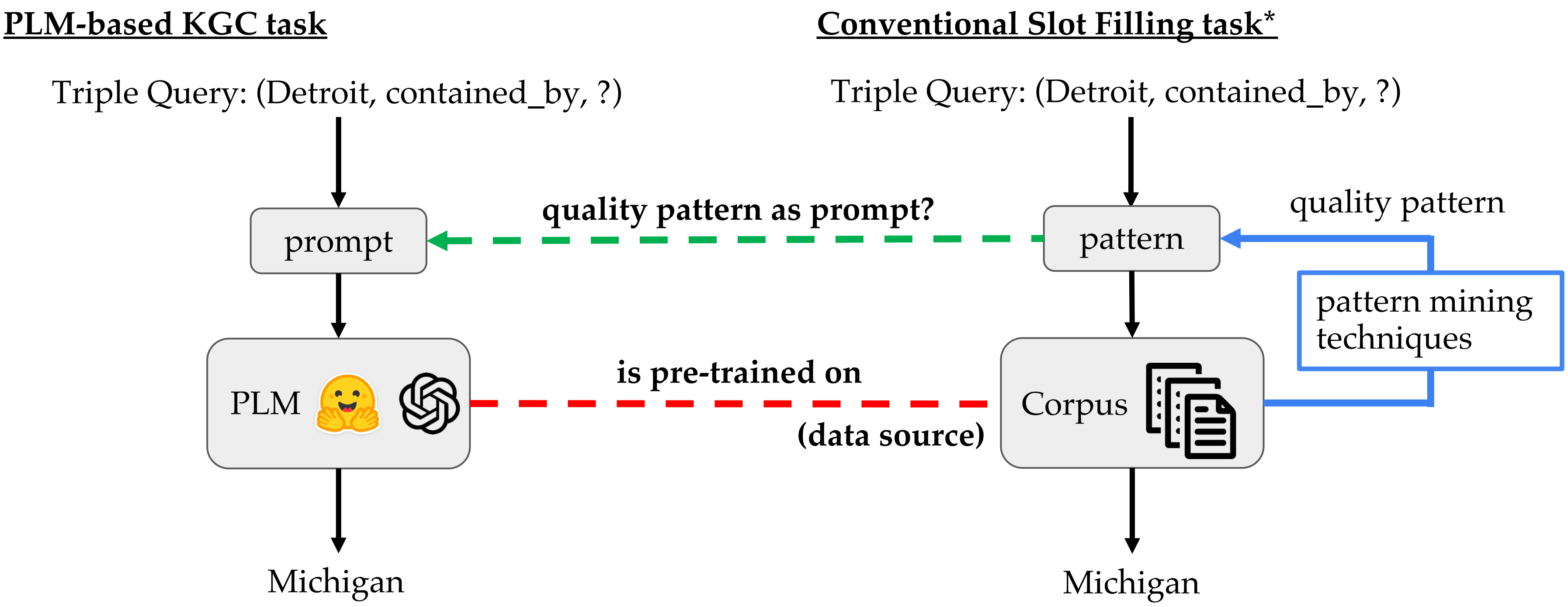
<u>Conventional Slot Filling task*</u>



* Yahya, Mohamed, et al. "Renoun: Fact extraction for nominal attributes." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Background & Motivation

How can we generate quality prompt automatically?



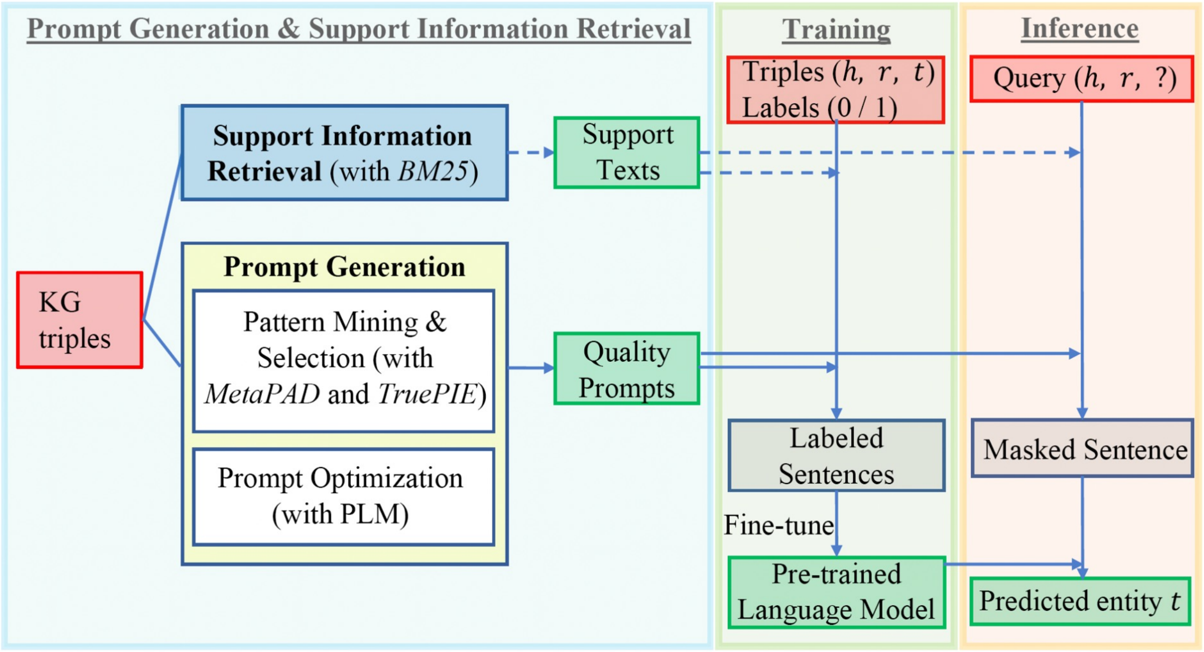
* Yahya, Mohamed, et al. "Renoun: Fact extraction for nominal attributes." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

Methodology

Methodology – TagReal

TagReal framework

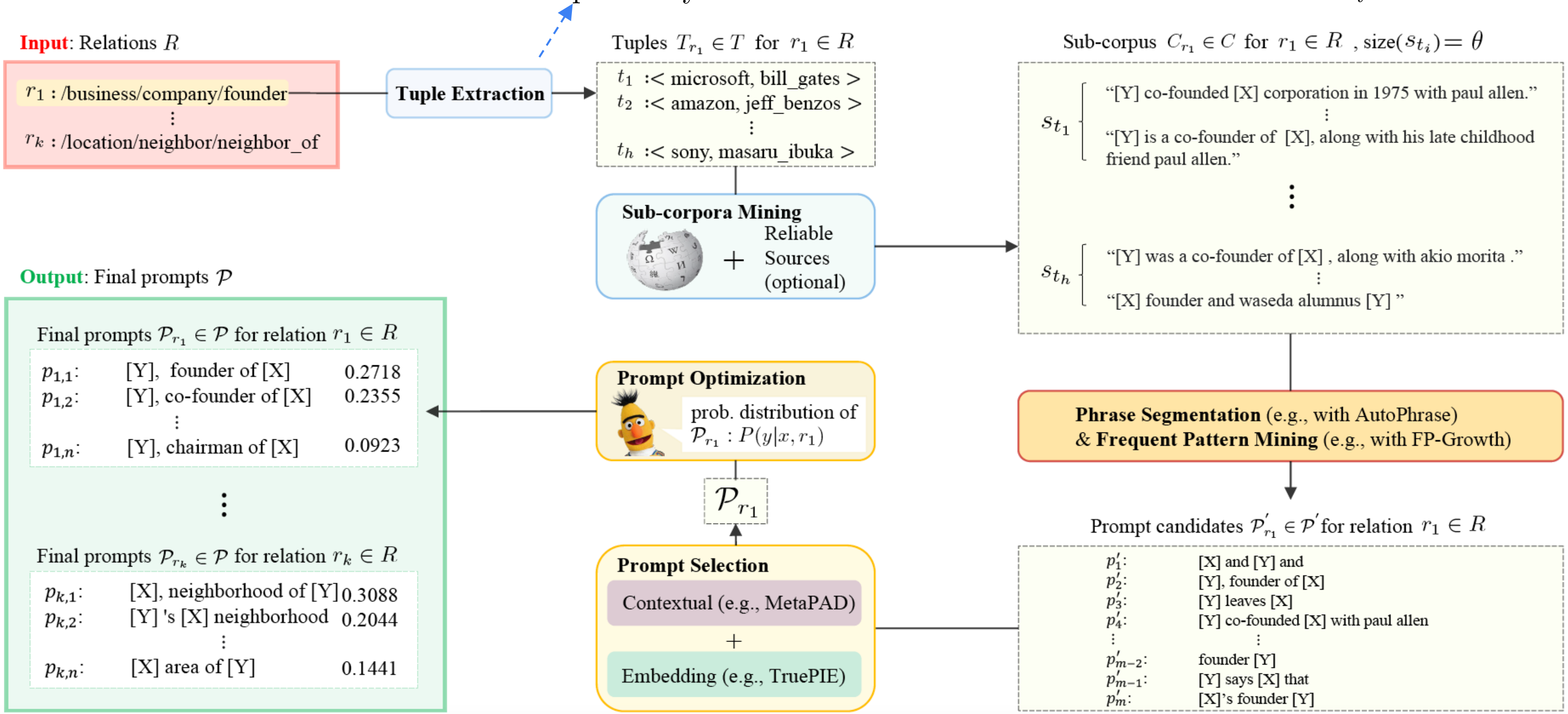
- Apply pattern mining & selection approaches to mine quality prompts from the corpus.
- (Optional) apply support information retrieval technique to retrieve relevant information from the corpus.



Methodology – TagReal

Prompt Generation Process

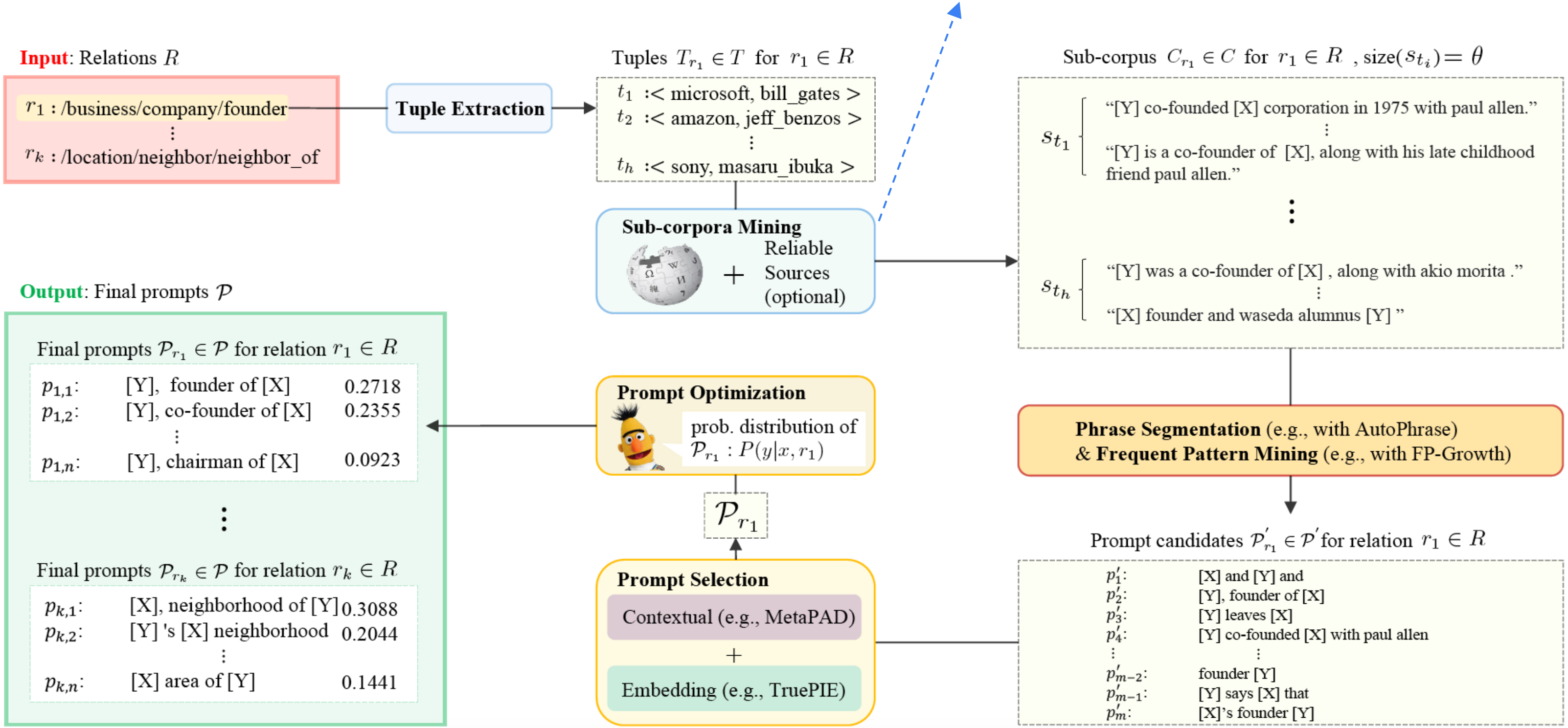
given a KG with a relation set $R = (r_1, r_2, ..., r_k)$, we first extract tuples T_{r_i} paired by head entities and tail entities for each relation $r_i \in R$ from the KG.



Methodology – TagReal

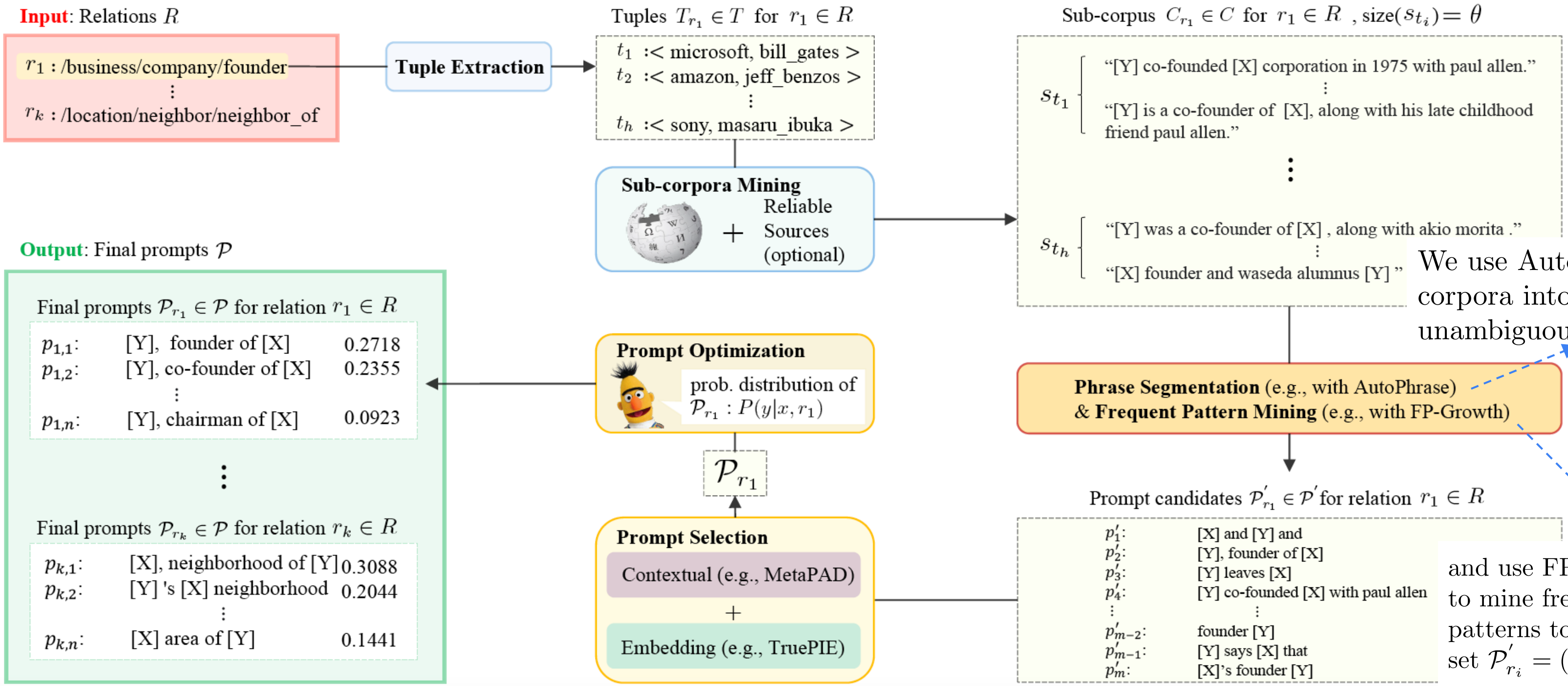
Prompt Generation Process

we then search sentences s_{t_j} containing both head and tail in a large corpus, to compose the sub-corpus C_{r_i}



Methodology – TagReal

Prompt Generation Process



We use AutoPhrase to segment corpora into more natural and unambiguous semantic phrases

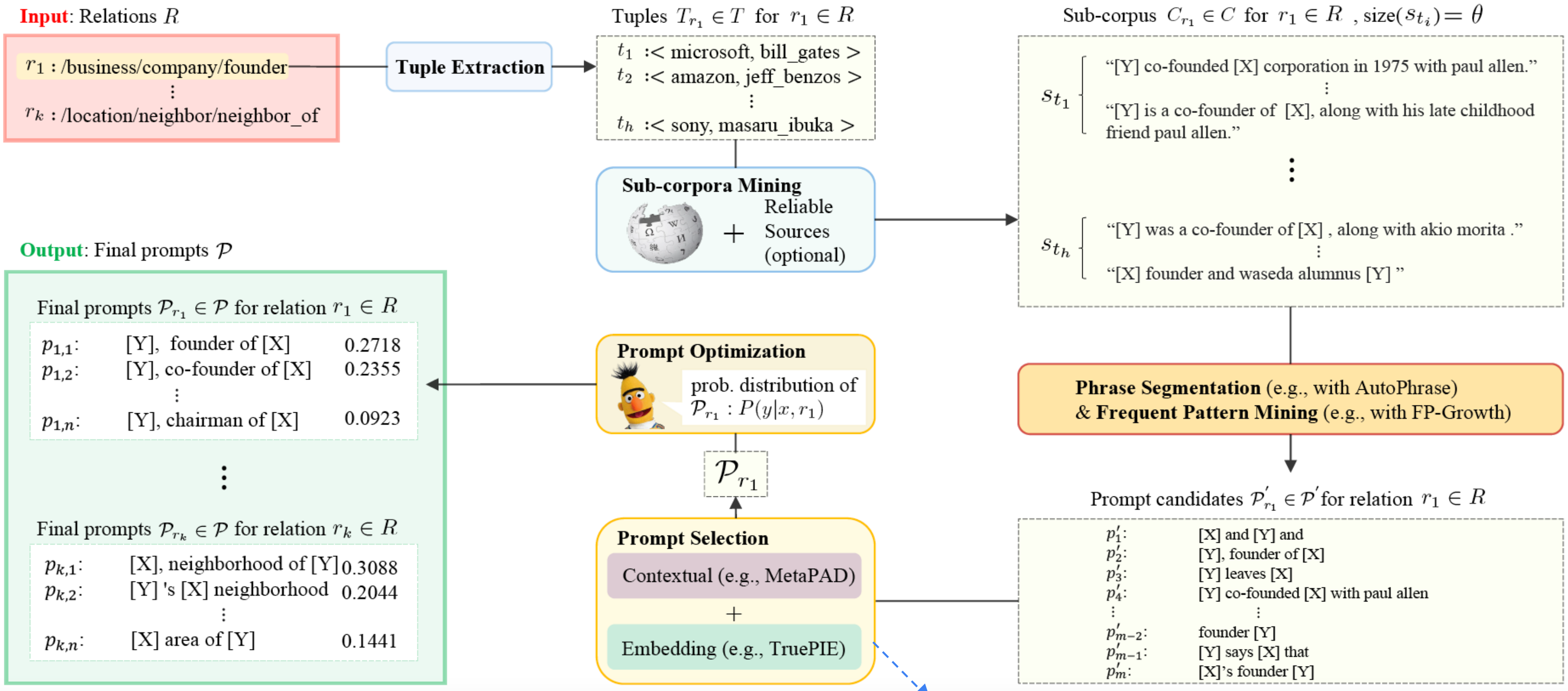
and use FP-Growth algorithm to mine frequent appeared patterns to compose a candidate set $\mathcal{P}'_{r_i} = (p'_1, p'_2, ..., p'_m)$.

(AutoPhrase) Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering, 30(10):1825–1837.

(FP-Growth) Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. ACM sigmod record, 29(2):1–12.

Methodology – TagReal

Prompt Generation Process

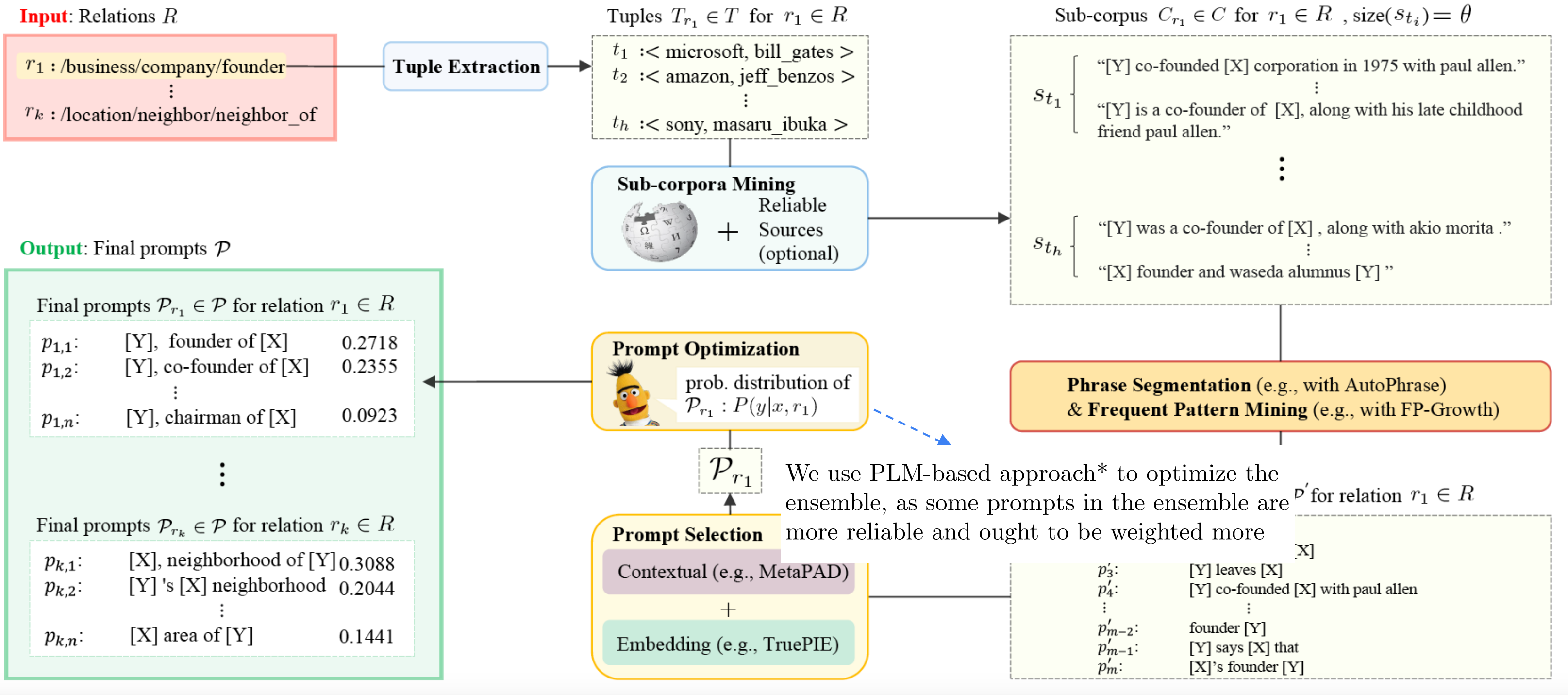


To select quality patterns from the candidate set, we apply two textual mining approaches: MetaPAD and TruePIE.

(MetaPAD) Jiang, Meng, et al. "Metapad: Meta pattern discovery from massive text corpora." *Proceedings of the 23rd ACM SIGKDD*. 2017.
(TruePIE) Li, Qi, et al. "Truepie: Discovering reliable patterns in pattern-based information extraction." *Proceedings of the 24th ACM SIGKDD*. 2018.

Methodology – TagReal

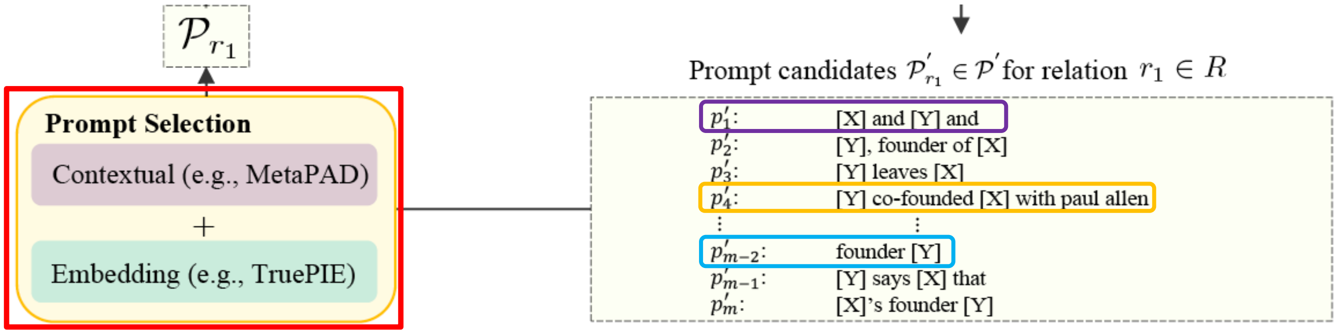
Prompt Generation Process



* Jiang, Zhengbao, et al. "How can we know what language models know?." *Transactions of the Association for Computational Linguistics* 8 (2020): 423-438.

Methodology – TagReal

The prompt selection is the most essential module in the prompt generation process.



Why it works?

MetaPAD

Metrics

(1) *Frequency*
(2) *Concordance*
(3) *Informativeness*
(4) *Completeness*
(5) *Coverage*

Since a PLM learns more contextual relations between frequent patterns and entities during the pre-training stage, a pattern occurs more frequently in the background corpus can probe more facts from the PLM.

if a pattern composed of highly associated sub-patterns appears frequently, it should be considered as a good one as the PLM would be familiar with the contextual relations among the sub-patterns.

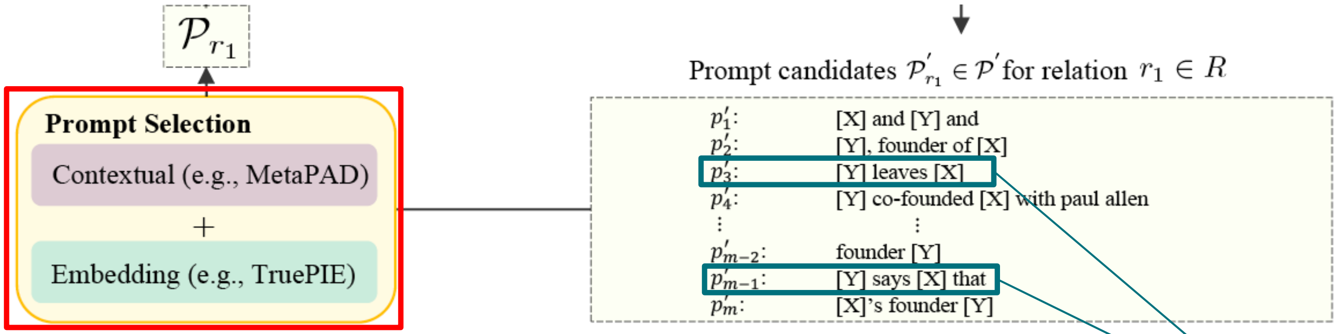
A pattern with low informativeness (e.g., p'_1) has the weak ability of PLM knowledge probing, as the relation between the subject or object entities cannot be well interpreted by it.

The completeness of a pattern affects a lot to the PLM knowledge probing especially when any of the placeholders is missing (e.g., p'_{m-2})

A quality pattern should be able to probe accurate facts from PLM as many as possible. Therefore, patterns like p'_4 which only suit a few or only one case should have a low quality score.

Methodology – TagReal

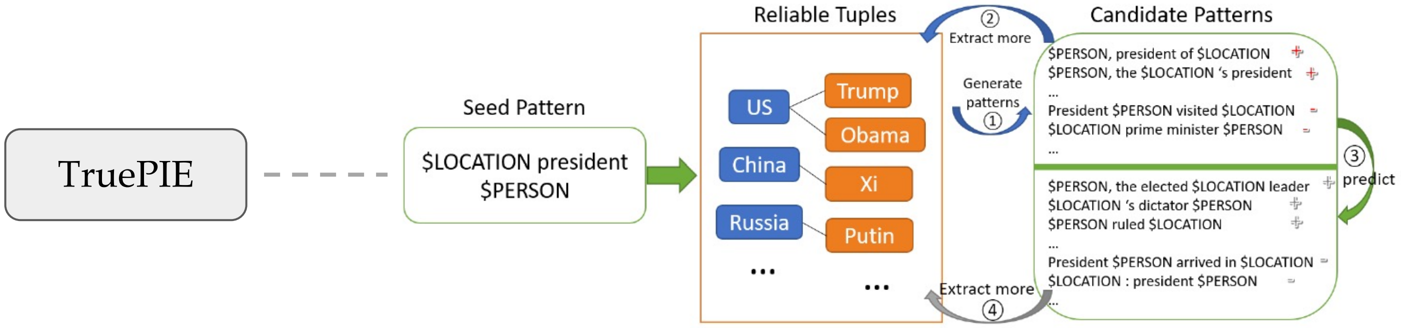
The prompt selection is the most essential module in the prompt generation process.



filtered

Why it works?

TruePIE filters the prompts that have low cosine similarity with the positive samples (e.g., p'_3 and p'_{m-1} are filtered), which matters to the creation of prompt ensemble since we want the prompts in the ensemble to be semantically close



Methodology – TagReal

Support Information Retrieval

For each triple, we search relevant sentences in corpus

Select the top-ranked sentence in length restriction

Attach the support texts to the prompts

(At the training phase, [MASK] is filled by object entity and [CLS] is filled by label)

Query q_i^r: < microsoft, /business/company/founder, ? >

↓ *BM25*

Support Information $S_{q_i^r}$

"however, microsoft is planning a significant marketing push into the field with a keynote speech by bill_gates, the company 's co-founder and chairman."

Prompt ensemble $\mathcal{P}_r$

p_1: [Y], founder of [X]
p_2: [Y], co-founder of [X]
⋮
p_n: [Y], chairman of [X]

↓

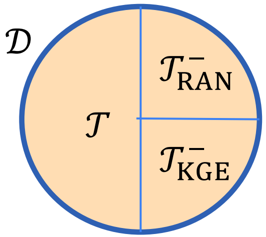
Query instances $\hat{q}_i^r$:

"[CLS] however, microsoft is planning a significant marketing push into the field with a keynote speech by bill_gates, the company 's co-founder and chairman. [SEP] [MASK], founder of microsoft"
⋮

(Negative Sampling)

$\mathcal{T}_{\text{RAN}}^-$: generated by randomly replacing the head or tail entity of the triple in $\mathcal{T}$ with other entity.

$\mathcal{T}_{\text{KGE}}^-$: generated by replacing the head or tail entity with another entity that KGE model considers to have a high probability of holding



$$|\mathcal{T}| = K \cdot |\mathcal{T}^-|$$

$$\frac{\mathcal{T}_{\text{RAN}}^-}{\mathcal{T}_{\text{KGE}}^-} = \frac{\alpha}{1-\alpha}$$

Experiments

Experiment – Settings

Dataset:
FB60K-NYT10 & UMLS-PubMed
(KG with the associated corpus)

Metrics:
Hits@N, Mean Reciprocal Rank (MRR)

$$Hits@N = \sum_{i=1}^{Q} \frac{\mathrm{R}_{i,in}}{Q} \text{ and } \mathrm{R}_{i,in} = \begin{cases} 0, \mathrm{R}_i > N \\ 1, \mathrm{R}_i \leq N, \end{cases}$$

$$MRR = \sum_{i=1}^{Q} \frac{1}{Q\mathrm{R}_i},$$

relations	#triples(all)	#queries(all)	ratio(all)	#triples(test)	#queries(test)	ratio(test)
FB60K-NYT10						
/people/person/nationality	44186	20215	2.19	4438	2282	1.94
/location/location/contains	42306	11971	3.53	4244	2373	1.79
/people/person/place_lived	29160	12760	2.29	3094	2066	1.50
/people/person/place_of_birth	28108	16341	1.72	2882	2063	1.40
/people/deceased_person/place_of_death	6882	4349	1.58	678	518	1.31
/people/person/ethnicity	5956	2944	2.02	574	305	1.88
/people/ethnicity/people	5956	2944	2.02	592	318	1.86
/business/person/company	4334	2370	1.83	450	379	1.19
/people/person/religion	3580	1688	2.12	300	175	1.71
/location/neighborhood/neighborhood_of	1275	547	2.33	130	91	1.43
/business/company/founders	904	709	1.28	94	87	1.08
/people/person/children	821	711	1.15	56	56	1.00
/location/administrative_division/country	829	498	1.66	88	72	1.22
/location/country/administrative_divisions	829	498	1.66	102	79	1.29
/business/company/place_founded	754	548	1.38	80	73	1.10
/location/us_county/county_seat	264	262	1.01	32	32	1.00
UMLS-PubMed						
may_be_treated_by	71424	7703	9.27	7020	3118	2.25
may_treat	71424	7703	9.27	6956	3091	2.25
may_be_prevented_by	10052	3232	3.11	1014	584	1.74
may_prevent	10052	3232	3.11	1034	586	1.76
gene_mapped_to_disease	6164	1732	3.56	596	331	1.80
disease_mapped_to_gene	6164	1732	3.56	652	357	1.82
gene_associated_with_disease	536	289	1.85	58	49	1.18
disease_has_associated_gene	536	289	1.85	48	41	1.17

Two datasets provided by Fu, et al.

* Fu, Cong, et al. "Collaborative Policy Learning for Open Knowledge Graph Reasoning." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Experiment – Baseline Comparison on KG Completion

Observation 1:
Non-PLM-based models suffer from training data dropping.

Reasons:
(1) KGE methods need very dense data to be trained well.
(2) Path finding-based methods like CPL are unable to recognize the underlying patterns with insufficient evidential and general paths.

Observation 2:
TagReal significantly outperforms the SOTA PLM-based method.

	Model	20%			50%			100%		
		Hits@5	Hits@10	MRR	Hits@5	Hits@10	MRR	Hits@5	Hits@10	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

Table 1: **Performance comparison of KG completion on FB60K-NYT10 dataset**. Results are averaged values of ten independent runs of head/tail entity predictions. The highest score is highlighted in **bold.**

	Model	20%		40%		70%		100%	
		Hits@5	Hits@10	Hits@5	Hits@10	Hits@5	Hits@10	Hits@5	Hits@10
KGE-based	TransE (Bordes et al., 2013)	19.70	30.47	27.72	41.99	34.62	49.29	40.83	53.62
	DisMult (Yang et al., 2014)	19.02	28.35	28.28	40.48	32.66	47.01	39.53	53.82
	ComplEx (Trouillon et al., 2016a)	11.28	17.17	24.64	35.15	25.89	38.19	34.54	49.30
	ConvE (Dettmers et al., 2018)	20.45	30.72	27.90	42.49	30.67	45.91	29.85	45.68
	TuckER (Balažević et al., 2019)	19.94	30.82	25.79	41.00	26.48	42.48	30.22	45.33
	RotatE (Sun et al., 2019)	17.95	27.55	27.35	40.68	34.81	48.81	40.15	53.82
Text&KGE-based	RC-Net (Xu et al., 2014)	7.94	10.77	7.56	11.43	8.31	11.81	9.26	12.00
	TransE+Line (Fu et al., 2019)	23.63	31.85	24.86	38.58	25.43	34.88	22.31	33.65
	JointNRE (Han et al., 2018)	21.05	31.37	27.96	40.10	30.87	44.47	-	-
RL-based	MINERVA (Das et al., 2017)	11.55	19.87	24.65	35.71	35.80	46.26	57.63	63.83
	CPL (Fu et al., 2019)	15.32	24.22	26.96	38.03	37.23	47.60	58.10	**65.16**
PLM-based	PKGC (Lv et al., 2022)	31.08	43.49	41.34	52.44	47.39	55.52	55.05	59.43
	TagReal (our method)	**35.83**	**46.45**	**46.26**	**55.99**	**53.46**	**60.40**	**60.68**	62.88

Table 2: **Performance comparison of KG completion on UMLS-PubMed dataset**. Results are averaged values of ten independent runs of head/tail entity predictions. The highest score is highlighted in **bold.**

Experiment – Ablation Study

Condition	FB60K-NYT10			UMLS-PubMed			
	20%	50%	100%	20%	40%	70%	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)

Table 3: **Ablation study on prompt and support information**. Data in brackets denotes Hits@5 (left) and Hits@10 (right). "man", "mine" and "optim" denote TAGREAL with manual prompts, mined prompt ensemble without optimization and optimized prompt ensemble, respectively. "supp" denotes application of support information.

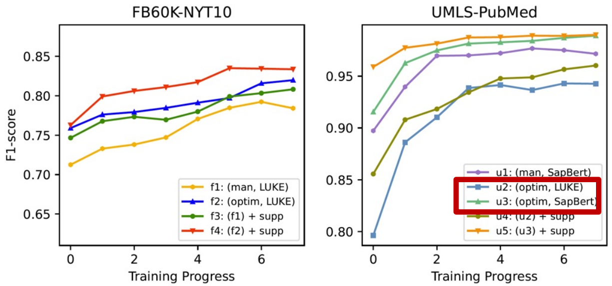


Figure 5: **Performance (F1-Score) variation of triple classification w.r.t training time.** "man" or "optim" means TAGREAL with manual prompts or optimized prompt ensemble. "supp" denotes support information.

Observation 1:
Support information retrieval helps, especially on the FB60K-NYT10 dataset.

Observation 2:
The ensemble of mined prompts can already outperform human-designed prompts.

Observation 3:
Weighted ensemble through PLM-based prompt optimization helps boost performance.

Observation 4:
The choice of PLM is important, especially for domain-specific datasets.

Experiment – Case Study

Query: (?, /location/location/contains, alba)

Manual Prompt	Optimized Prompt Ensemble	weights
[Y] is located in [X].	[Y], [X] .	0.10490836
	home in [Y], [X] .	0.23949857
	[Y] is in [X] .	0.24573646
	school in [Y], [X] .	0.32810964
	people from [Y], [X] .	0.34946583

Support Information (retrieved by BM25)

"**in alba** , **italy** 's truffle capital , **in the northwestern province of piedmont** , demand for the fungi has spawned a cottage industry of package tours , food festivals and a strip mall of truffle-themed shops . "

Predictions (Top10 in descending order of classification scores)

Man :	united_states_of_america, pennsylvania, france, lombardy, abruzzo, jamaica, **piedmont**, ivrea, massachusetts, iraq
Optim :	cuneo, **piedmont**, italy, sicily, lazio, texas, campania, northern_italy, scotland, calabria
Man + Supp :	sicily, italy, massachusetts, lazio, **piedmont**, united_states_of_america, abruzzo, tuscany, iraq, milan
Optim + Supp :	**piedmont**, cuneo, italy, northern_italy, canale, tuscany, campania, sicily, lazio, calabria

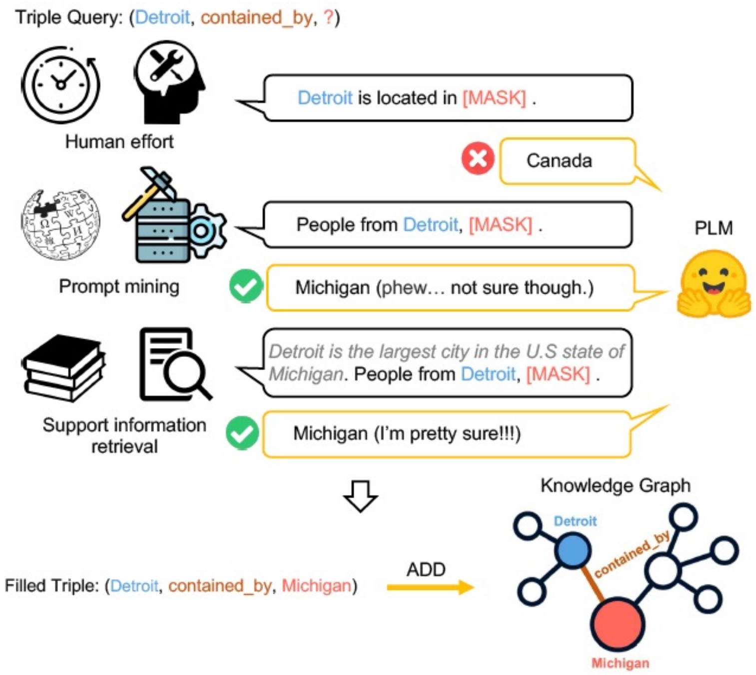
Figure 7: **Example of the link prediction with TagReal on FB60K-NYT10**. **Man** denotes manual prompt. **Optim** denotes optimized prompt ensemble. **Supp** denotes support information. The **ground truth tail entity**, **helpful information** and **optimized prompts** (darker for higher weights) are highlighted.

Observation :
Support information helps, but is not as essential as optimized prompts.

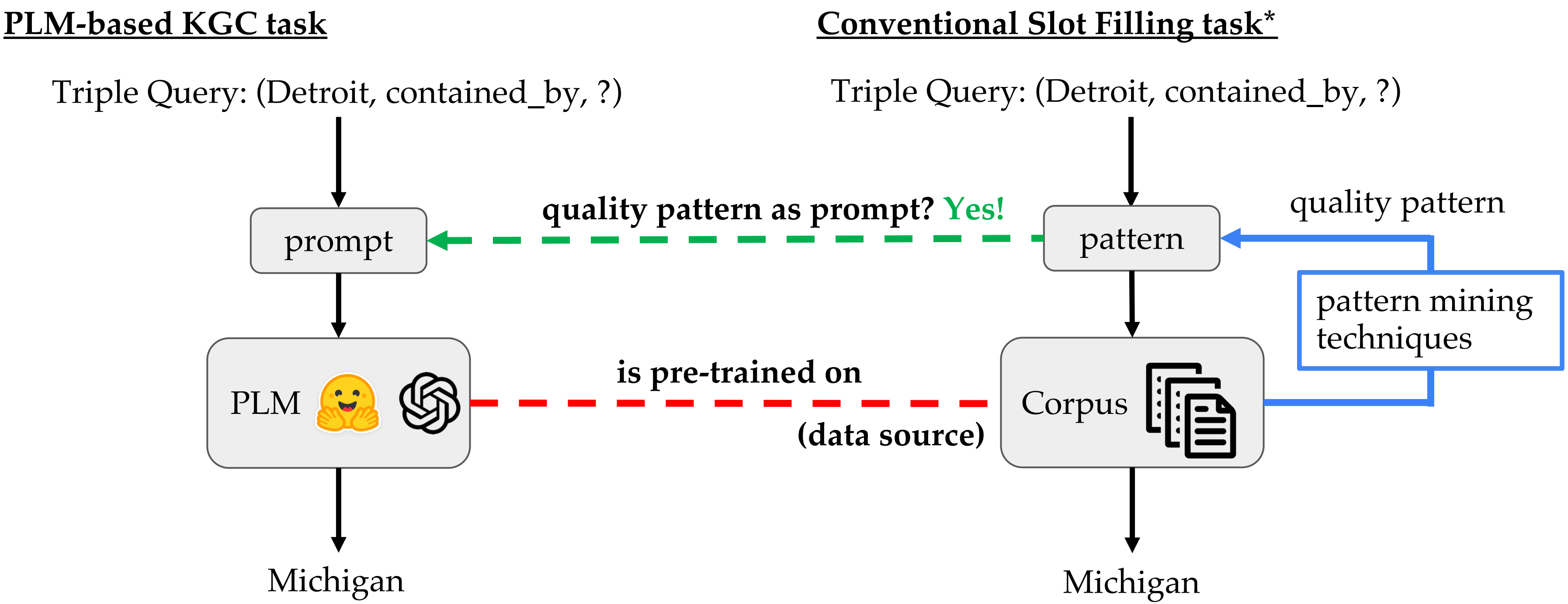
Conclusion & Thoughts

Conclusion

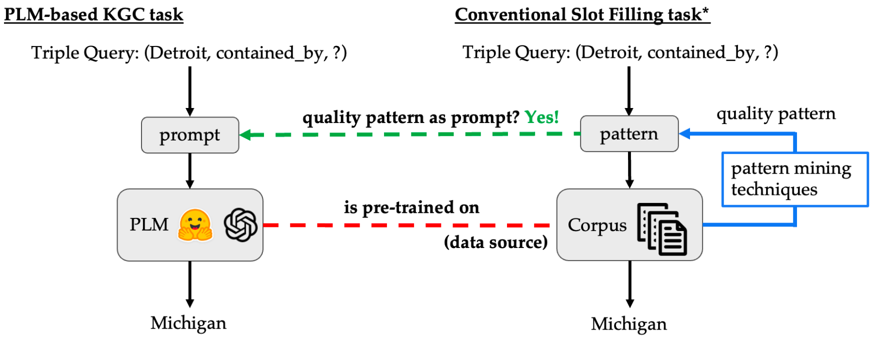
- We proposed a novel framework that combines quality prompt generation and support information retrieval, to exploit implicit knowledge in PLM for the knowledge graph completion task.
- Experimental results show that our method could perform much better than previous non-PLM-based methods especially when the training data is limited.
- We demonstrated that the prompts generated by our approach are better than the human-designed ones. The support information retrieval could also boost performance.



Thoughts


PLM-based KGC task
Conventional Slot Filling task*
Triple Query: (Detroit, contained_by, ?)
Triple Query: (Detroit, contained_by, ?)
prompt
quality pattern as prompt? Yes!
pattern
quality pattern
pattern mining techniques
PLM
is pre-trained on
(data source)
Corpus
Michigan
Michigan

Thoughts



- We believe this is the seed work bridging the gap between traditional text/pattern mining and contemporary prompt mining methods.
- This work is a step forward in our understanding of how text mining methods could provide a new avenue to analyze the workings of pre-trained language models.

Thanks for your attention!

Our code is available on:
https://github.com/pat-jj/TagReal
Further questions?
- Email me (Pengcheng (Patrick) Jiang) at pj20@illinois.edu