


CoLAKE: Contextualized Language and Knowledge Embedding

Tianxiang Sun, Yunfan Shao, Xipeng Qiu
Qipeng Guo, Yaru Hu, Xuanjing Huang, Zheng Zhang

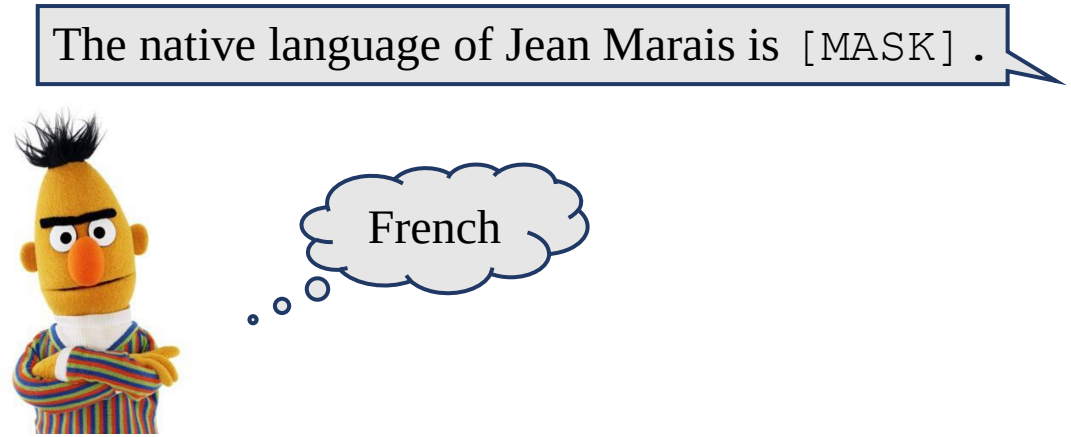
COLING 2020

Language Models Need Knowledge

- **PLMs perform poorly on entity recognition**
  - Contextualized PLMs achieved small improvements on entity & semantic related tasks compared with non-contextualized methods. (Tenney et al.)
  - BPE tokenization breaks entities

    `The native language of Jean Mara ##is is French.`
  - Surface form-based reasoning

	original BERT	E-BERT-replace	E-BERT-concat	ERNIE	Know-Bert
Jean Marais	French	French	French	french	french
Daniel Ceccaldi	Italian	French	French	french	italian
Orane Demazis	Albanian	French	French	french	french
Sylvia Lopez	Spanish	French	Spanish	spanish	spanish
Annick Alane	English	French	French	english	english

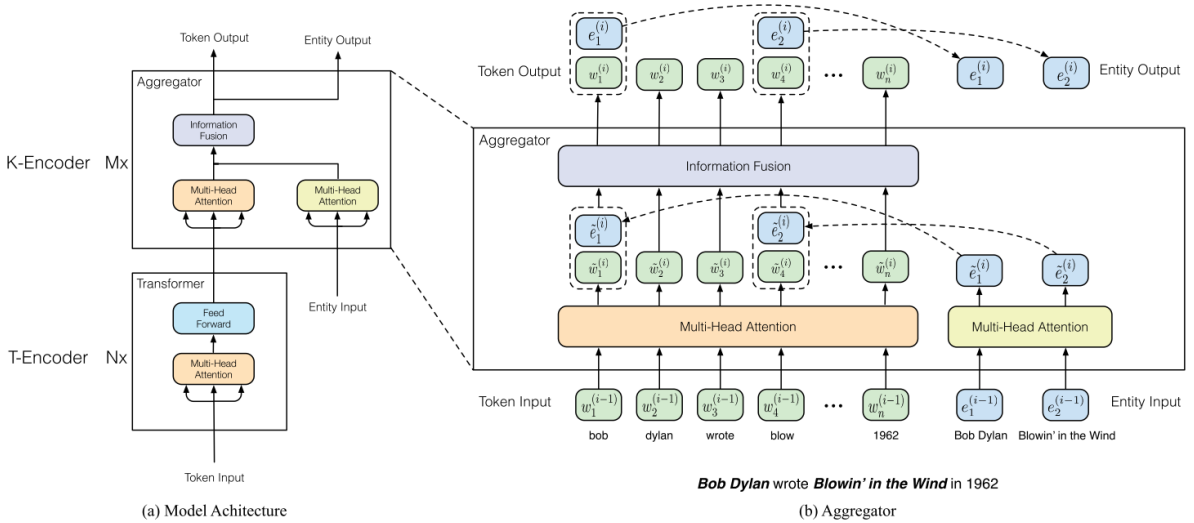


BERT does not know *Daniel Ceccaldi* (as an entity) at all.
It just think *Daniel Ceccaldi* looks like an Italian name.

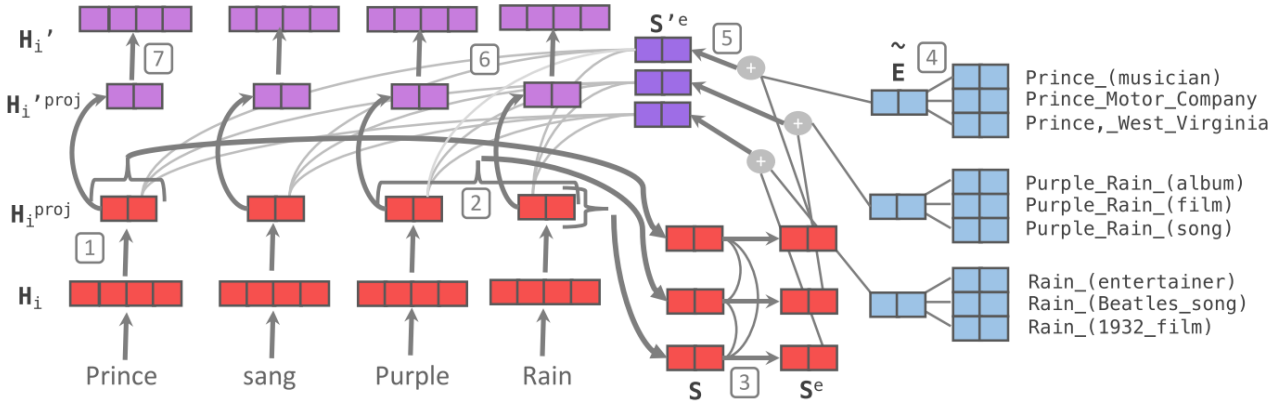
E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. https://arxiv.org/abs/1911.03681

Injecting Knowledge into PLMs

- **Injecting entity embeddings**
  - ERNIE, KnowBERT, K-BERT, etc.
  - The entity embeddings are NOT
    - Jointly learned along with PLM
    - Contextualized
- **Knowledge as supervision**
  - WKLM, etc.



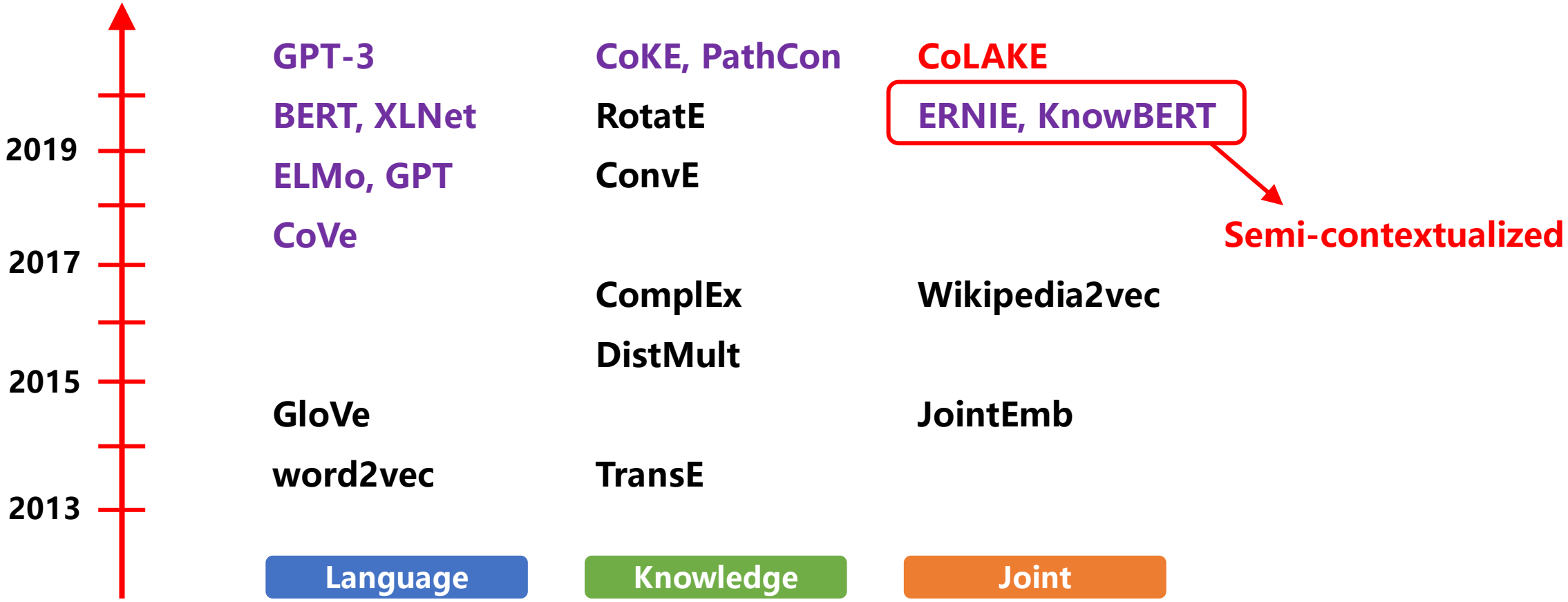
ERNIE: Enhanced Language Representation with Informative Entities
https://arxiv.org/abs/1905.07129



Knowledge Enhanced Contextual Word Representations
https://arxiv.org/abs/1909.04164

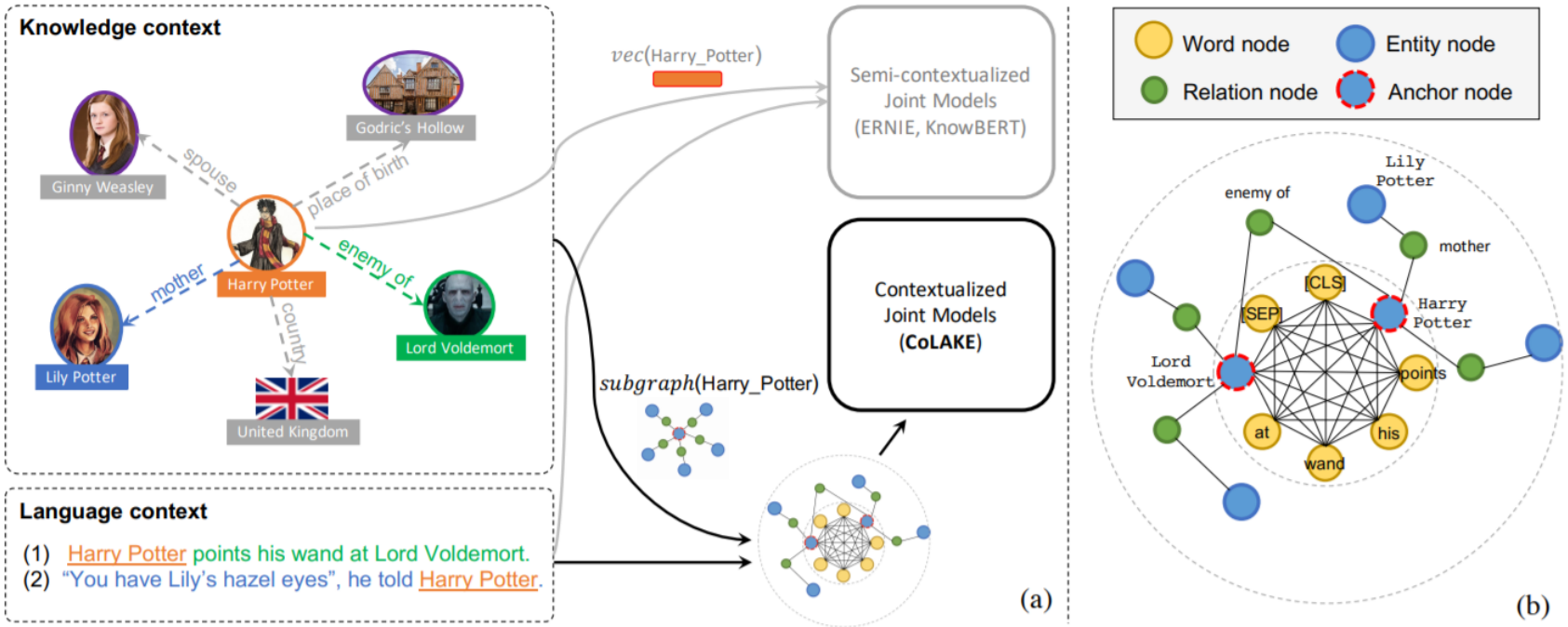
Representation in Language and Knowledge

- Combine the success of both sides -- CoLAKE



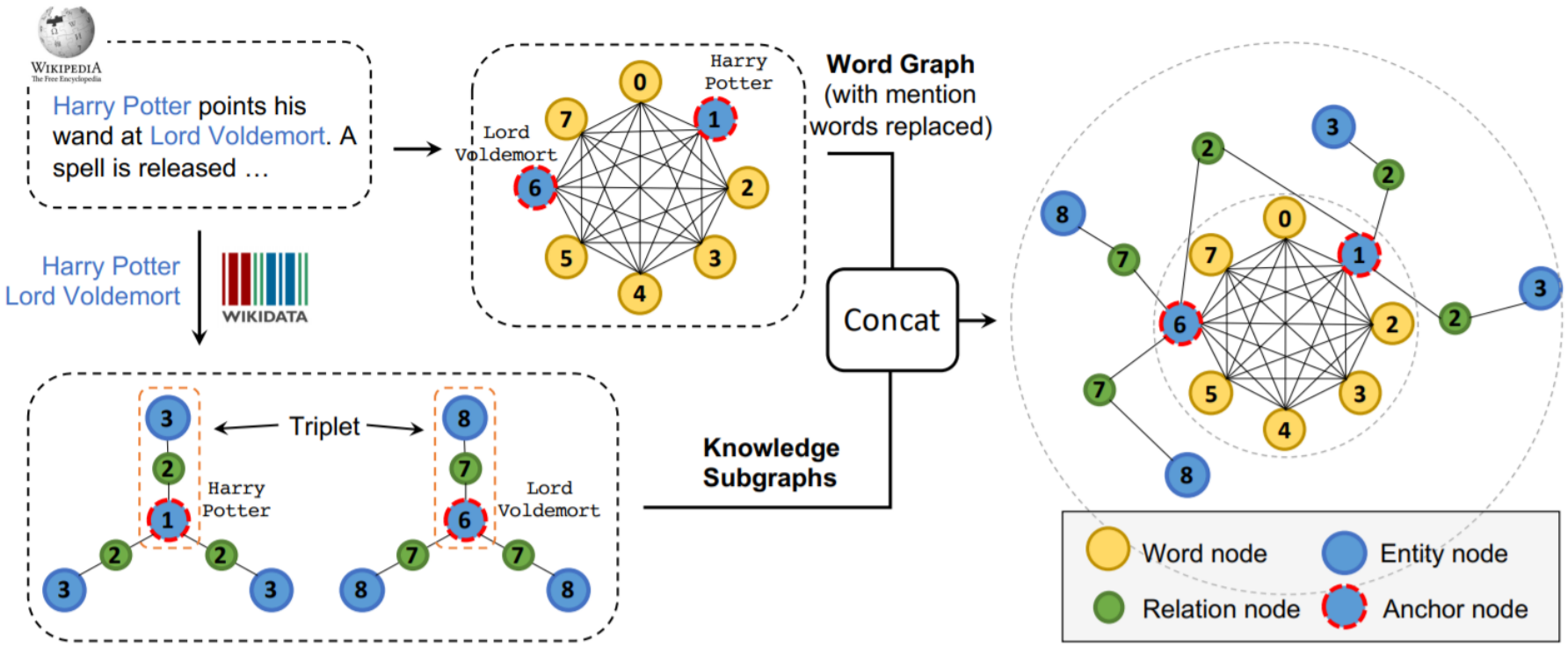
Word-knowledge Graph

- **Why?**
  - Different facts should be accessed to help understand different sentences.
- **What?**
  - Word-knowledge graph is a unified data structure to integrate language context and knowledge context.



Word-knowledge Graph

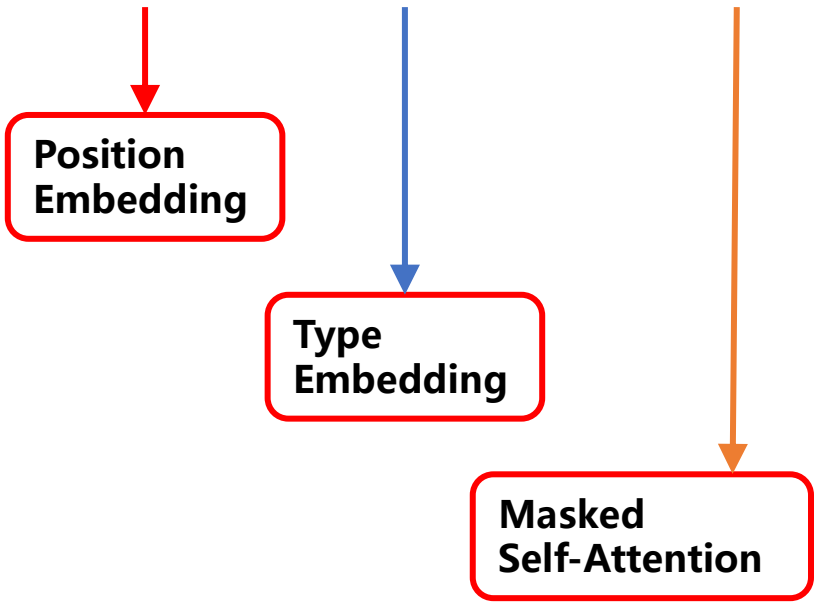
- **How?**
  - Word graph + Knowledge subgraph

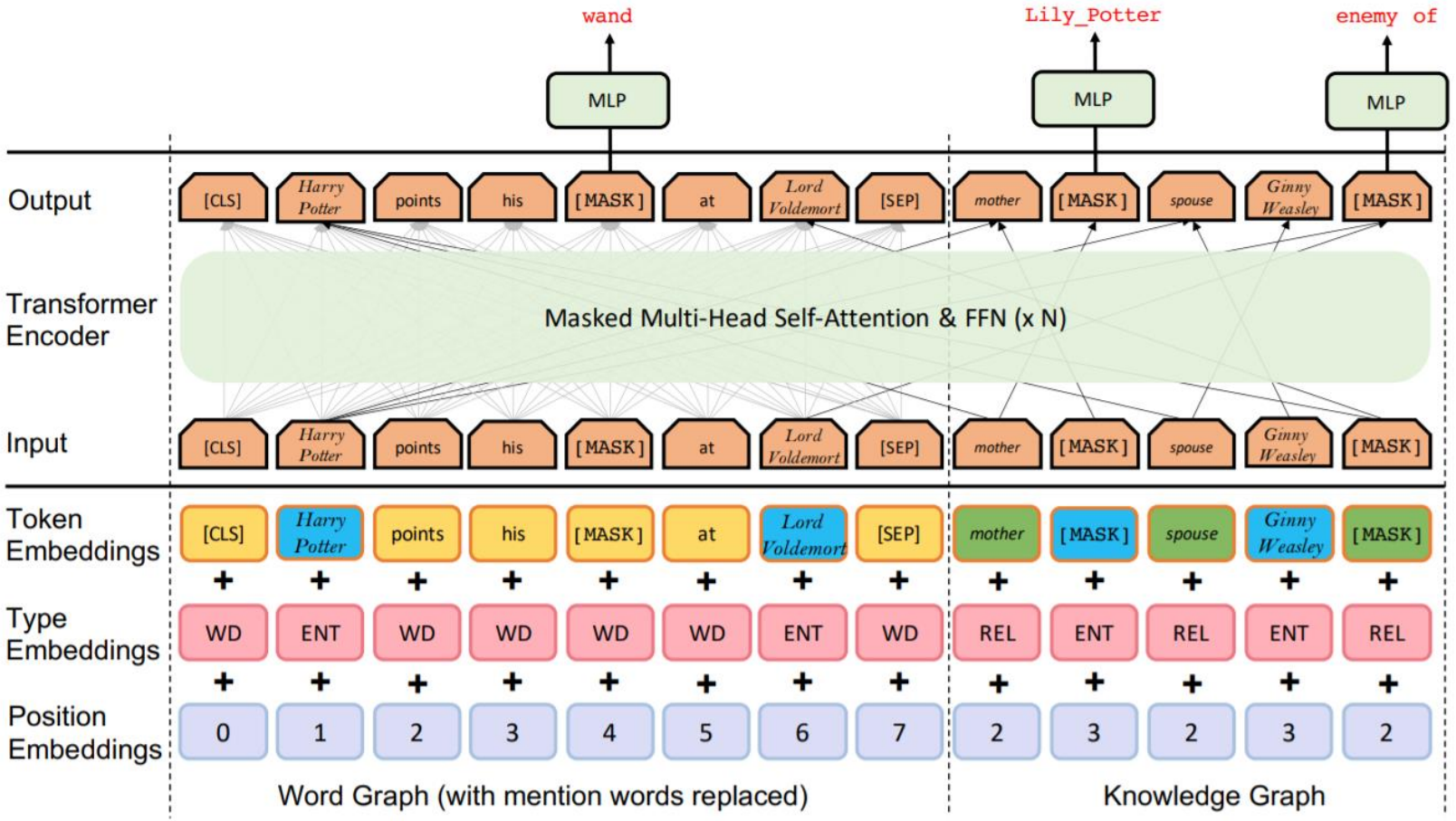


Model Architecture

- **Modify Transformer for word-knowledge graph**

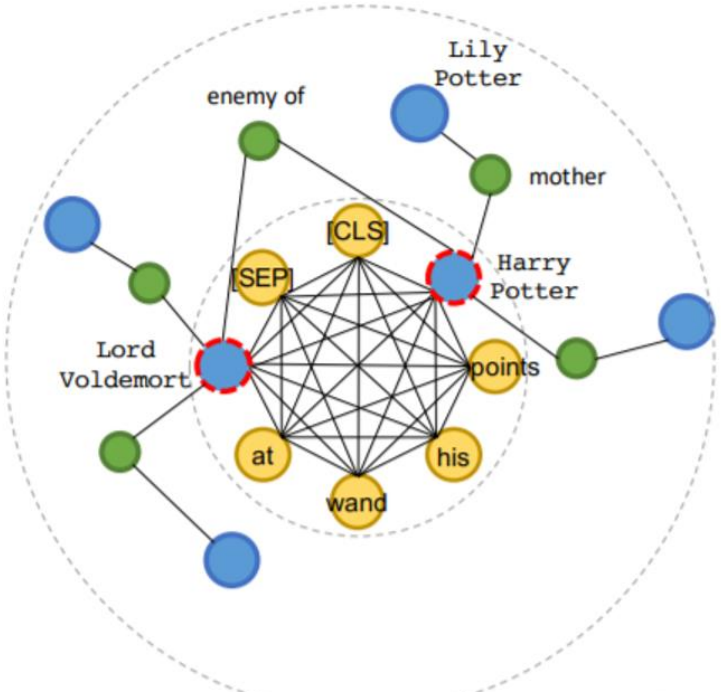
Word-knowledge graph is a positional heterogeneous graph.





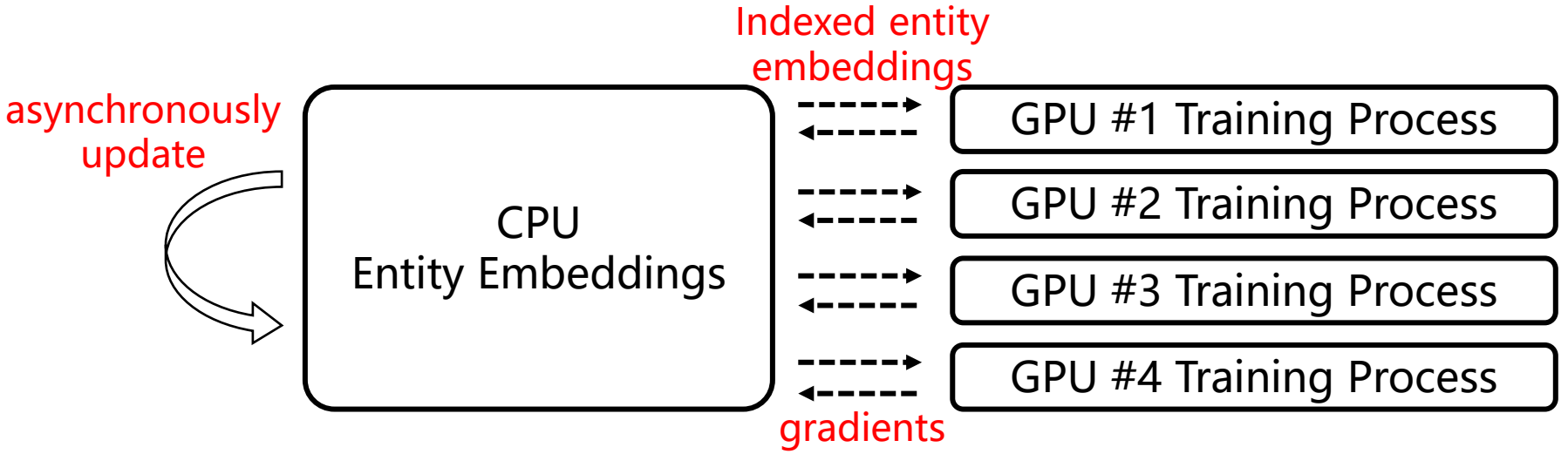
Pre-Training Objective

- **Masked Language Model (MLM) on word-knowledge graph**
  - Masking word nodes
    - Learn linguistic knowledge
  - Masking entity nodes
    - Anchor nodes masked – Learn to align the two spaces
    - Entity nodes masked – Learn contextualized entity embeddings
  - Masking relation nodes
    - Relation between anchor nodes – Learn to do relation extraction
    - Otherwise – Learn contextualized relation embeddings



Some Details

- **Mixed CPU-GPU training**



- **Negative sampling**
  - Sample negative entities from the 3/4 powered entity distribution.
- **Alignment of the two spaces**
  - Discard neighbors of anchor nodes in 50% of time.
  - Replace mention words with anchor nodes.

Experiments

- **Knowledge-driven tasks**
  - Entity typing
  - Relation extraction
- **Knowledge probing tasks**
  - LAMA
  - LAMA-UHN
- **Language understanding tasks**
  - GLUE
- **Synthetic graph task**
  - Word-knowledge graph completion

Experiments

- **Knowledge-driven tasks**

Model	Open Entity			FewRel		
	P	R	F	P	R	F
BERT (Devlin et al., 2019)	76.4	71.0	73.6	85.0	85.1	84.9
RoBERTa (Liu et al., 2019)	77.4	73.6	75.4	85.4	85.4	85.3
ERNIE (Zhang et al., 2019)	78.4	72.9	75.6	88.5	88.4	88.3
KnowBERT (Peters et al., 2019)	**78.6**	73.7	76.1	-	-	-
KEPLER (Wang et al., 2019c)	77.8	74.6	76.2	-	-	-
E-BERT (Pörner et al., 2019)	-	-	-	88.6	88.5	88.5
CoLAKE (Ours)	77.0	**75.7**	**76.4**	**90.6**	**90.6**	**90.5**

- **Knowledge probing tasks**

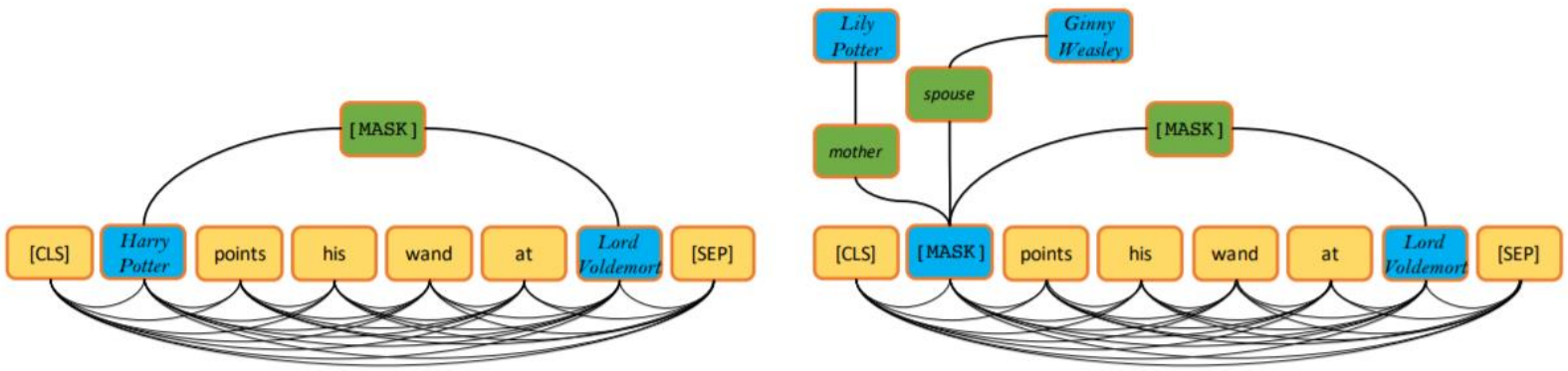
Corpus	Pre-trained Models					
	ELMo	ELMo5.5B	BERT	RoBERTa	CoLAKE	K-Adapter*
LAMA-Google-RE	2.2	3.1	11.4	5.3	9.5	7.0
LAMA-UHN-Google-RE	2.3	2.7	5.7	2.2	4.9	3.7
LAMA-T-REx	0.2	0.3	32.5	24.7	28.8	29.1
LAMA-UHN-T-REx	0.2	0.2	23.3	17.0	20.4	23.0

Experiments

- **Language understanding tasks**

Model	MNLI (m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	AVG.
RoBERTa	87.5 / 87.3	91.9	92.8	94.8	63.6	91.2	90.2	78.7	86.4
KEPLER	87.2 / 86.5	91.5	92.4	94.4	62.3	89.4	89.3	70.8	84.9
CoLAKE	87.4 / 87.2	92.0	92.4	94.6	63.4	90.8	90.9	77.9	86.3

- **Synthetic graph task**
  - Word-knowledge graph completion



(a) Transductive setting.

(b) Inductive setting.

Experiments

- **Results on word-knowledge graph completion**
  - CoLAKE is essentially a pre-trained inductive GNN which simultaneously models structural knowledge and text semantics.

Model	MR ↓	MRR	HITS@1	HITS@3	HITS@10
Transductive setting					
TransE (Bordes et al., 2013)	15.97	67.30	60.28	70.96	79.75
DistMult (Yang et al., 2015)	27.09	60.56	48.66	69.69	79.61
ComplEx (Trouillon et al., 2016)	26.73	61.09	49.80	70.64	79.78
RotatE (Sun et al., 2019)	30.36	70.90	64.74	74.89	81.05
CoLAKE	2.03	82.48	72.14	92.19	98.58
Inductive setting					
DKRL (Xie et al., 2016)	168.21	8.18	5.03	7.28	14.13
CoLAKE	31.01	28.10	15.69	30.28	58.05

Thanks!

CoLAKE: Contextualized Language and Knowledge Embedding, Sun et al. 2020

https://github.com/txsun1997/CoLAKE