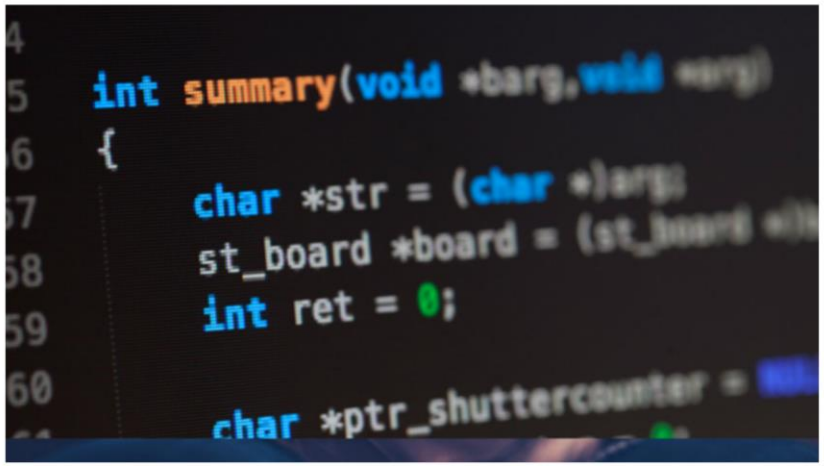
融合知识的预训练语言模型

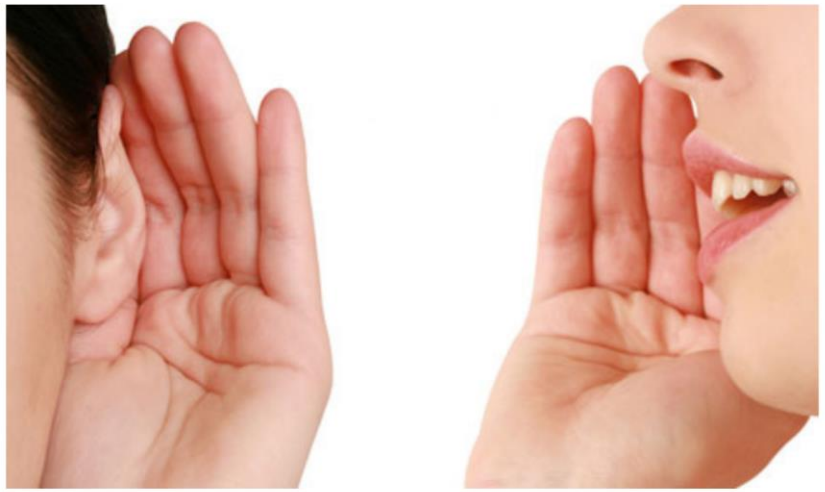
By im0qianqian

Background (自然语言处理)

- 自然语言是人类间交流传播信息和知识的工具
  - 创新性，歧义性，>CFG



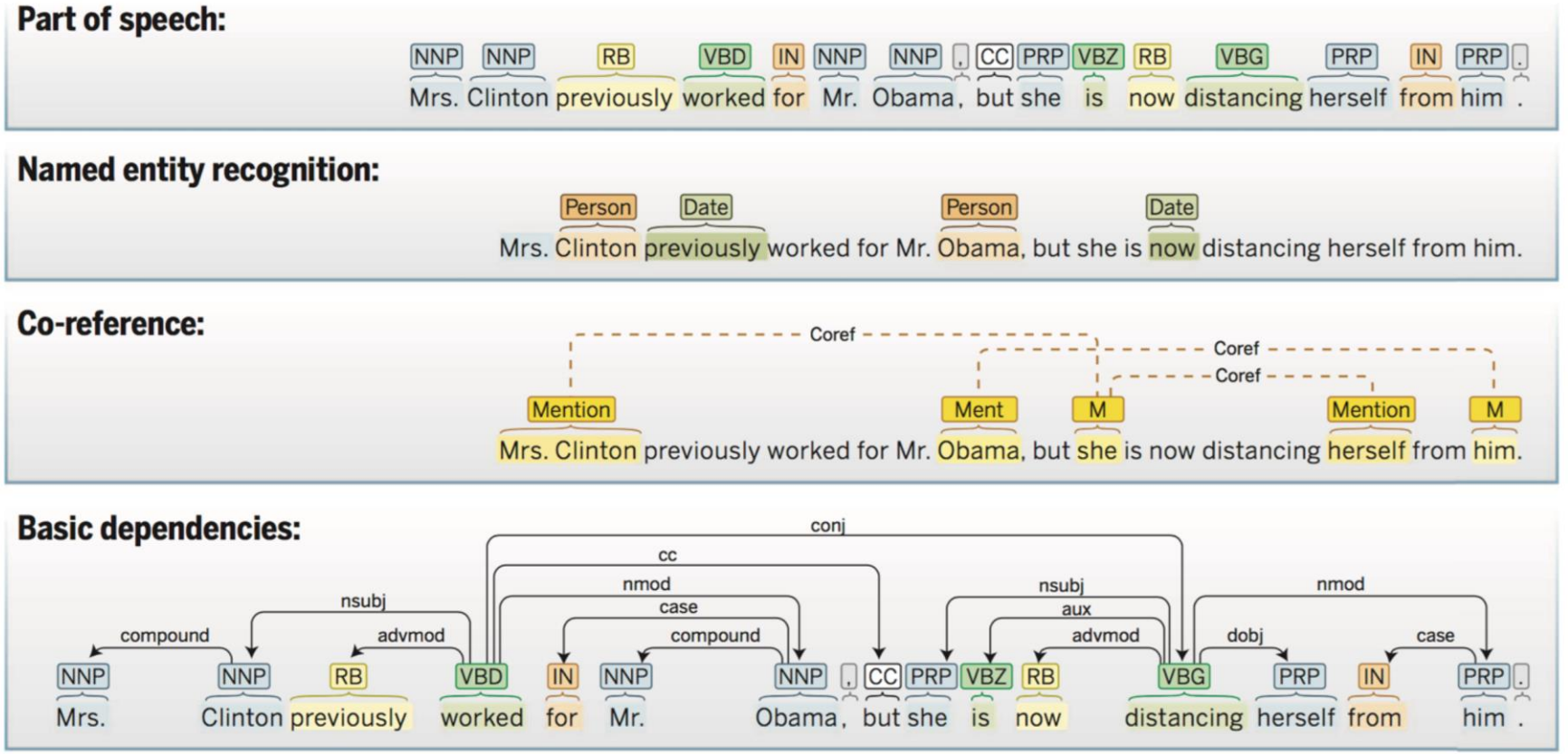
编程语言



自然语言

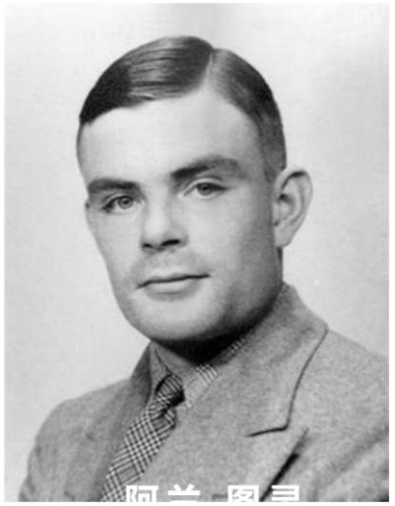
Background (自然语言处理)

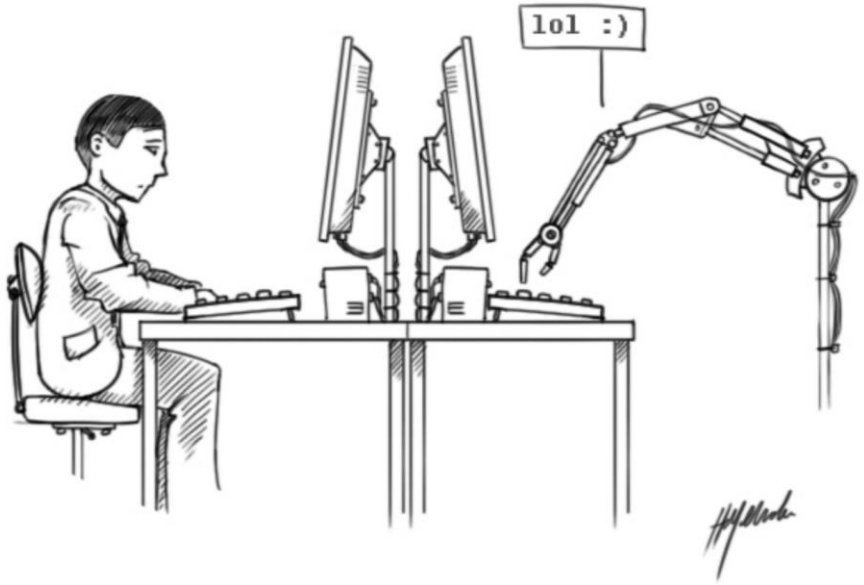
- 自然语言处理旨在理解人类语言的语义信息
- 本质是从无结构序列中预测有结构语义

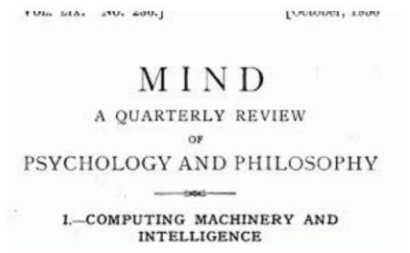


http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Background (自然语言处理)





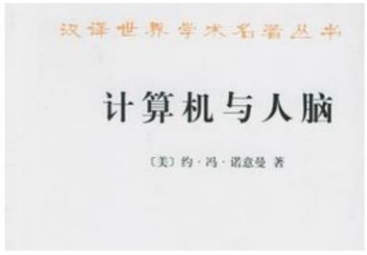

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

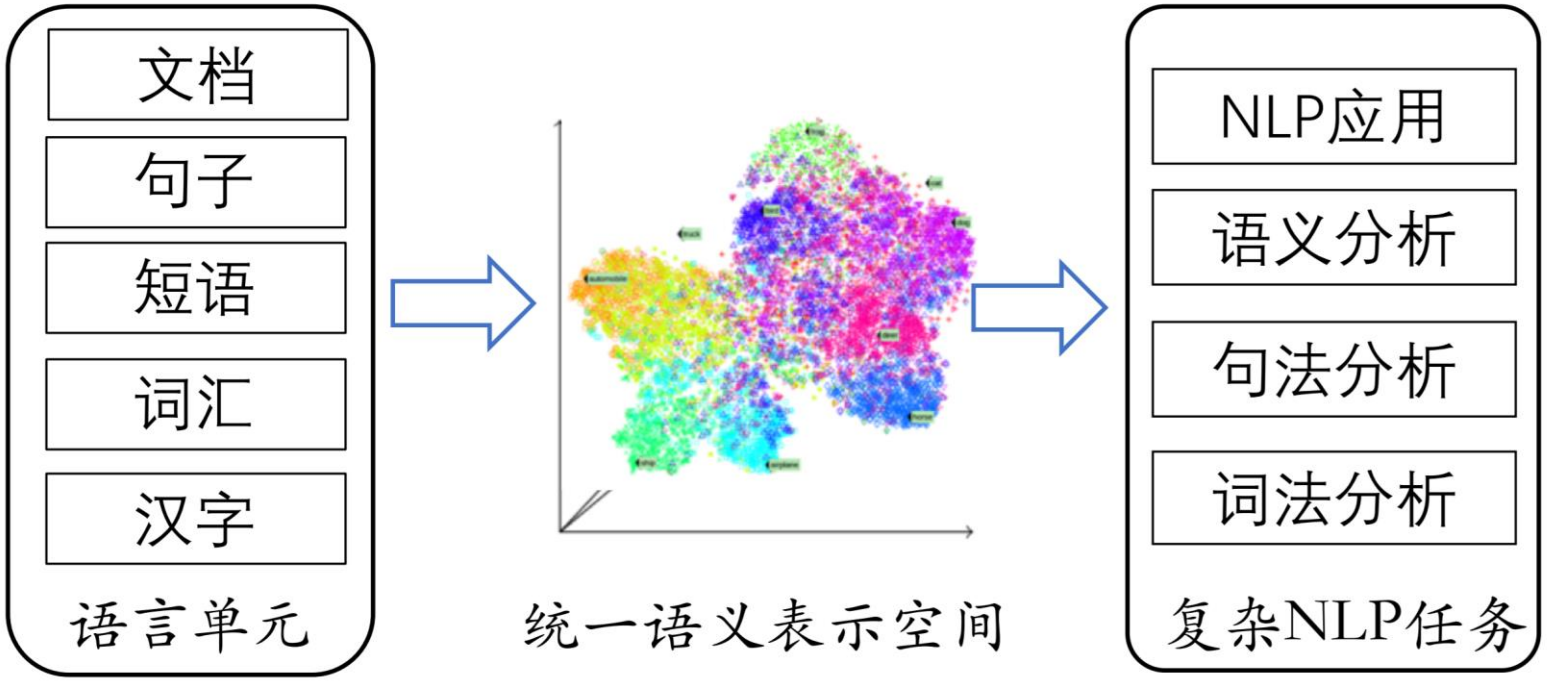


自然语言处理是实现人工智能、通过图灵测试的关键

http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Background (数据驱动下的自然语言处理)

- 深度学习技术在自然语言处理取得了巨大突破

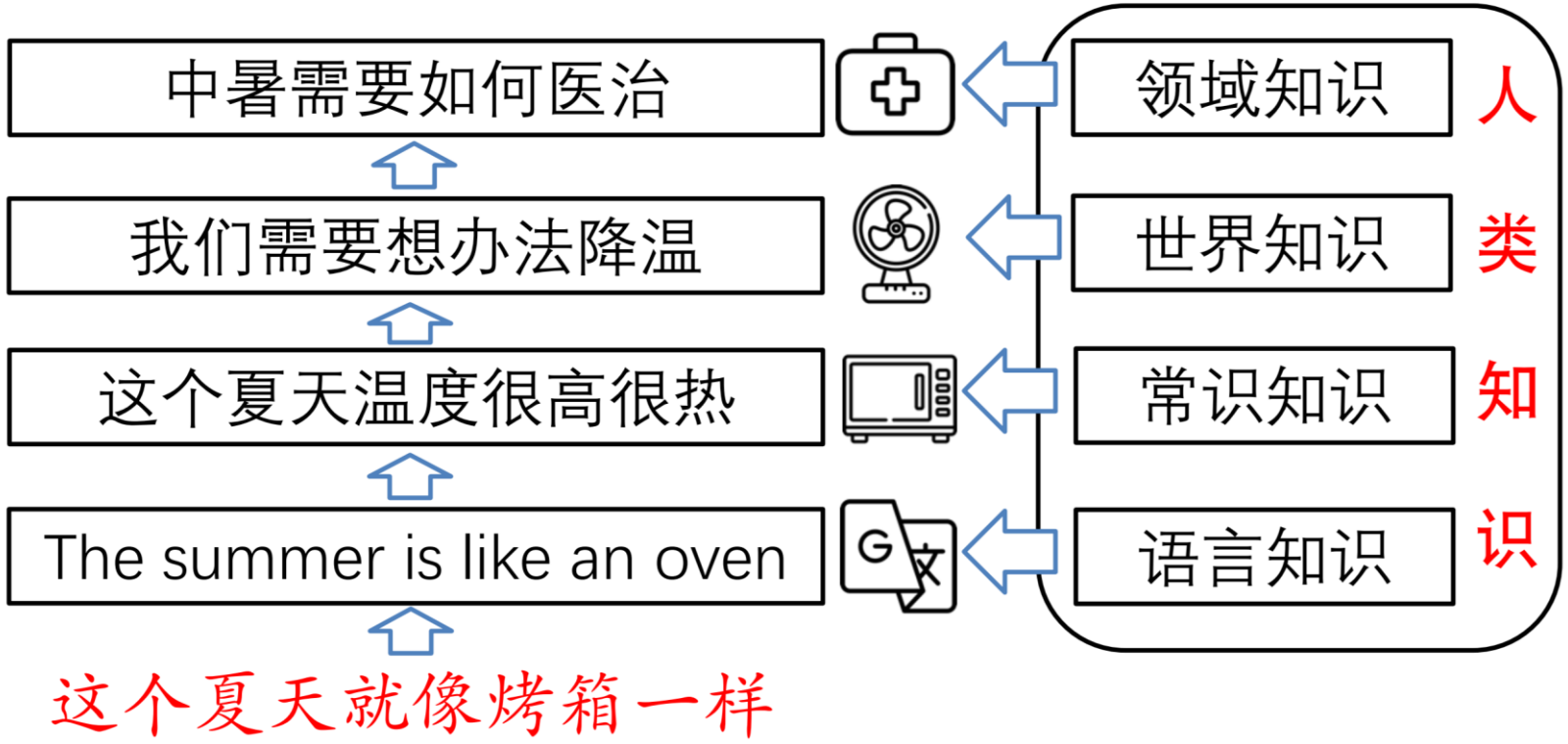


深度学习能够高效学习多粒度语言单元间复杂语义关联

http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Background (面临挑战)

- 对自然语言的深度理解需要复杂知识的支持

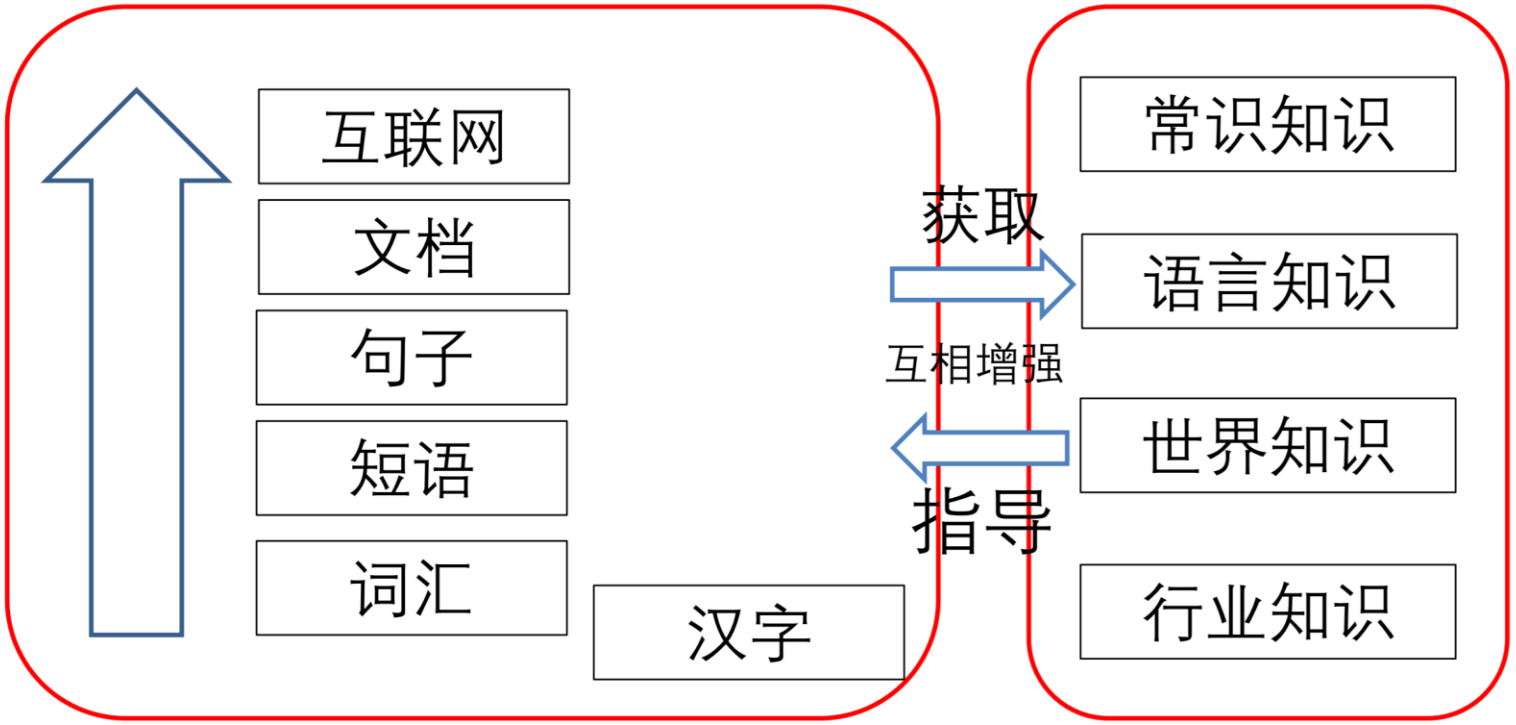


亟需知识支持实现NLP从字面意思到言外之意的跃迁

http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Background (自然语言特点)

- 自然语言文本蕴含丰富的语言知识和世界知识

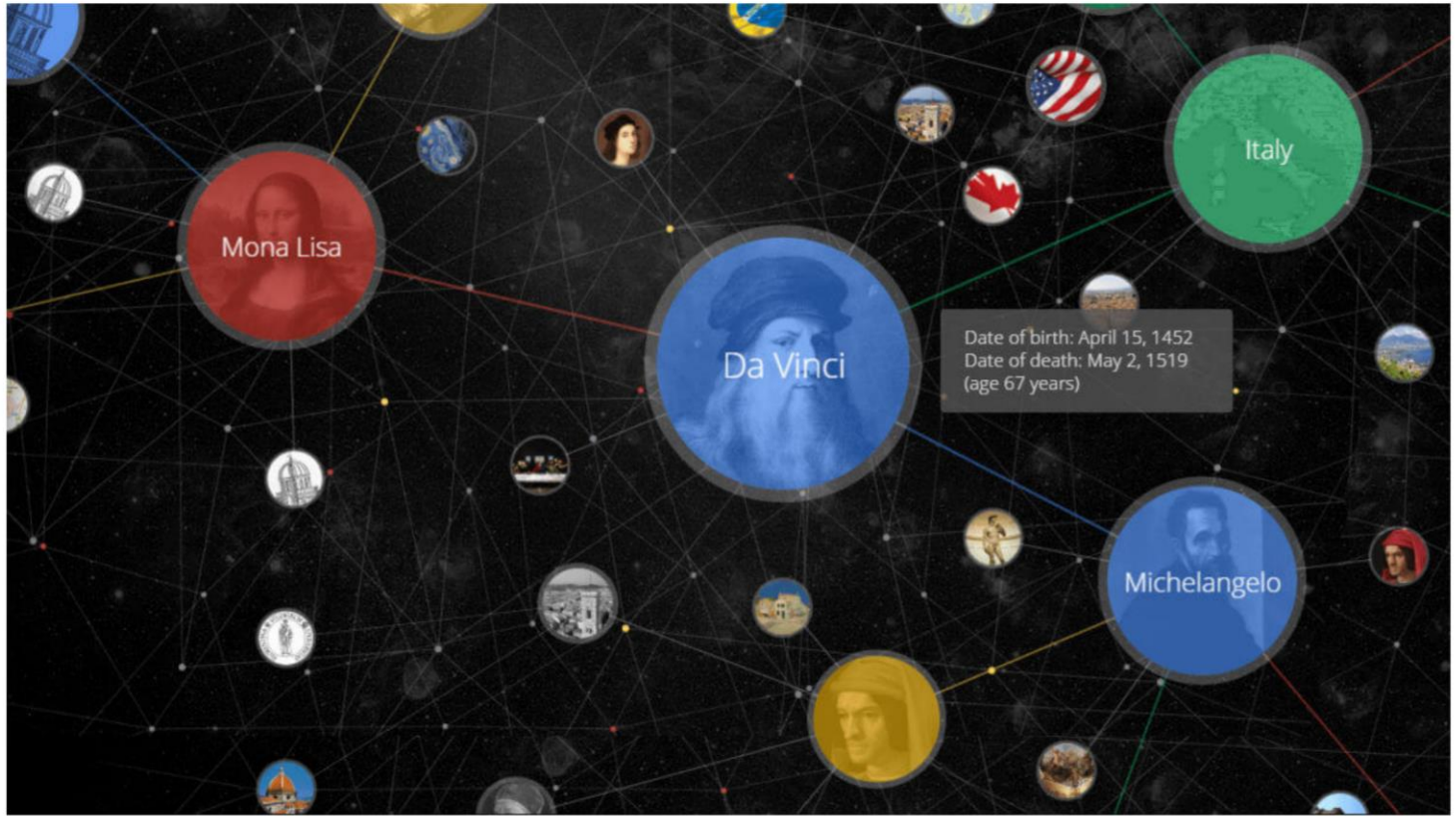


数据驱动 + 知识指导

http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Background (技术挑战)

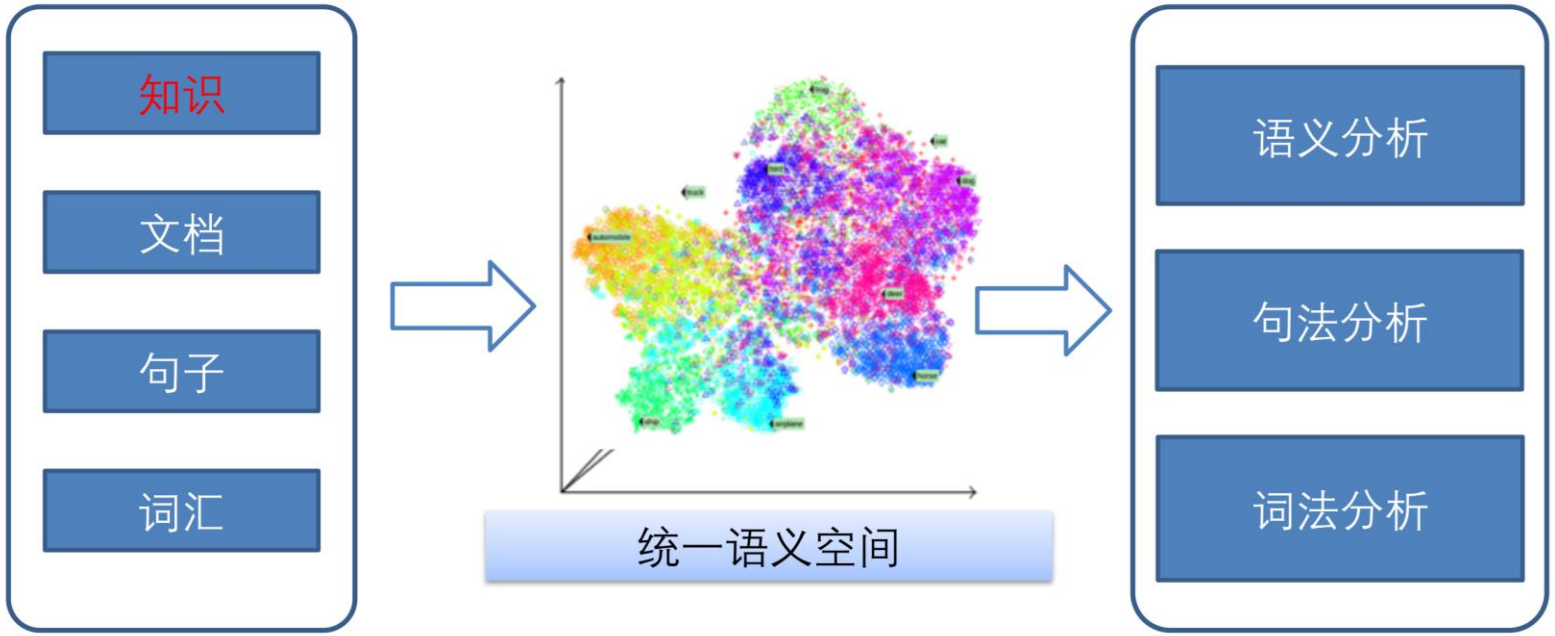
- 语言知识、世界知识均通过离散符号表示



http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

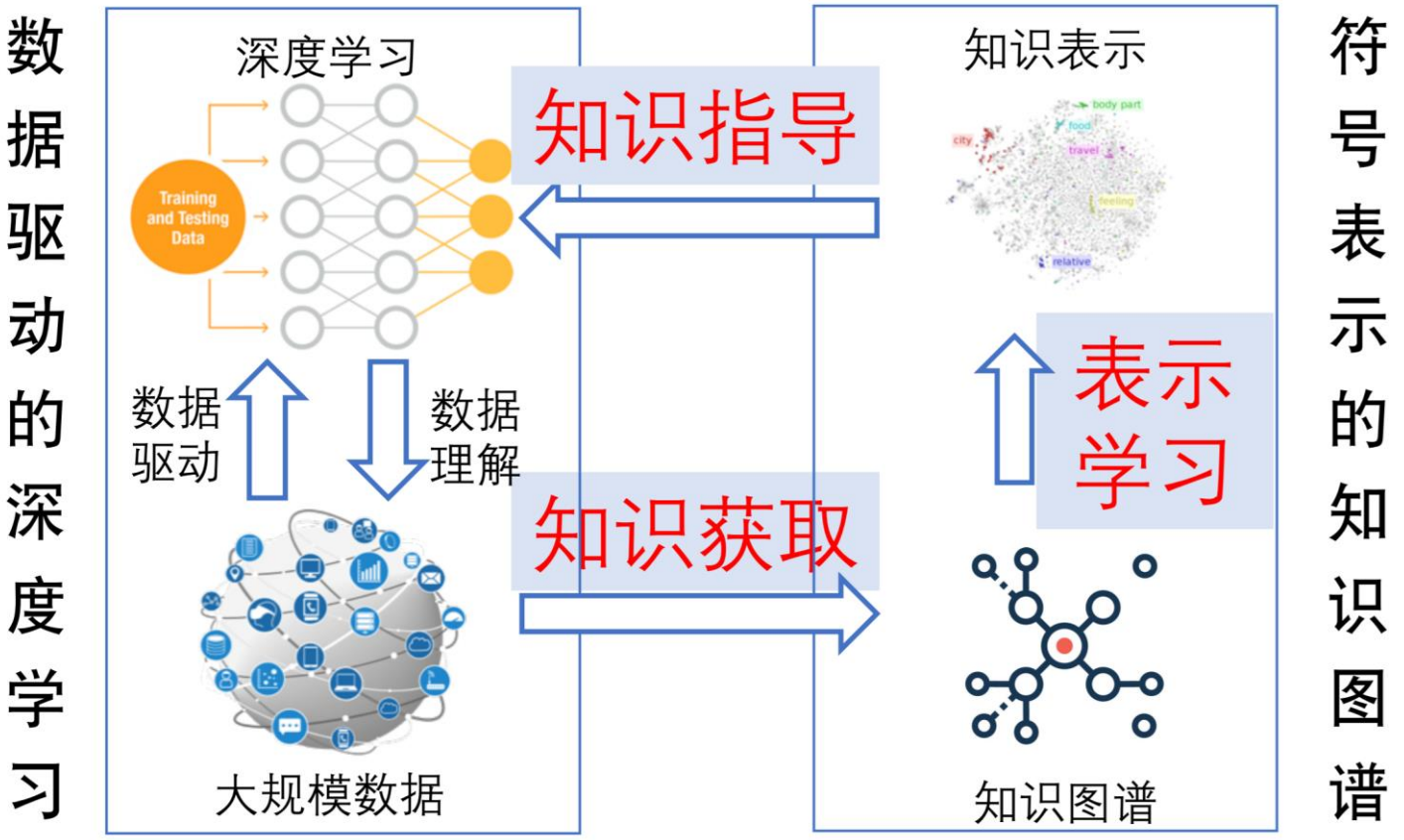
Background (表示学习)

- 分布式表示：实现跨粒度、跨领域、富知识的语言理解



http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Background (研究思路)

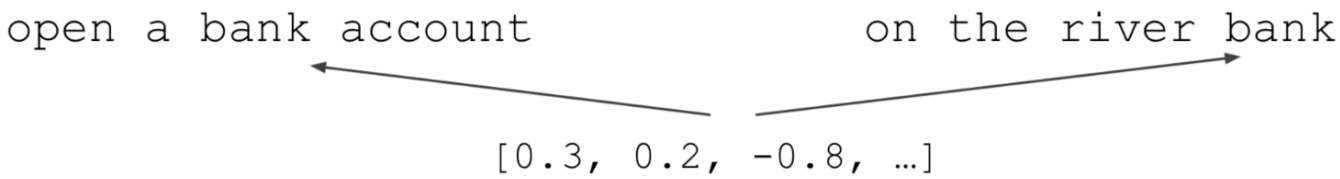


深度学习+知识图谱 双向驱动的自然语言处理技术体系

http://nlp.csai.tsinghua.edu.cn/~lzy/talks/2019_knowledge_guided_nlp_cn.pdf

Feature-based PLMs

- ELMO
  - 早期的 Word2Vec、Glove 被广泛用于各类 NLP 任务
    - 但是信息单一，词汇在不同语境下的复杂语义难以体现

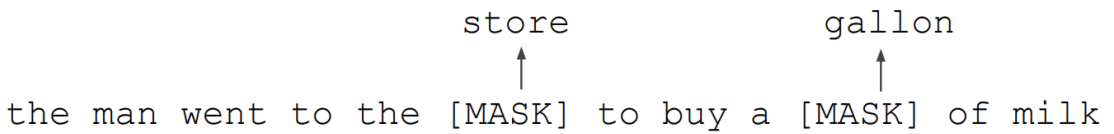


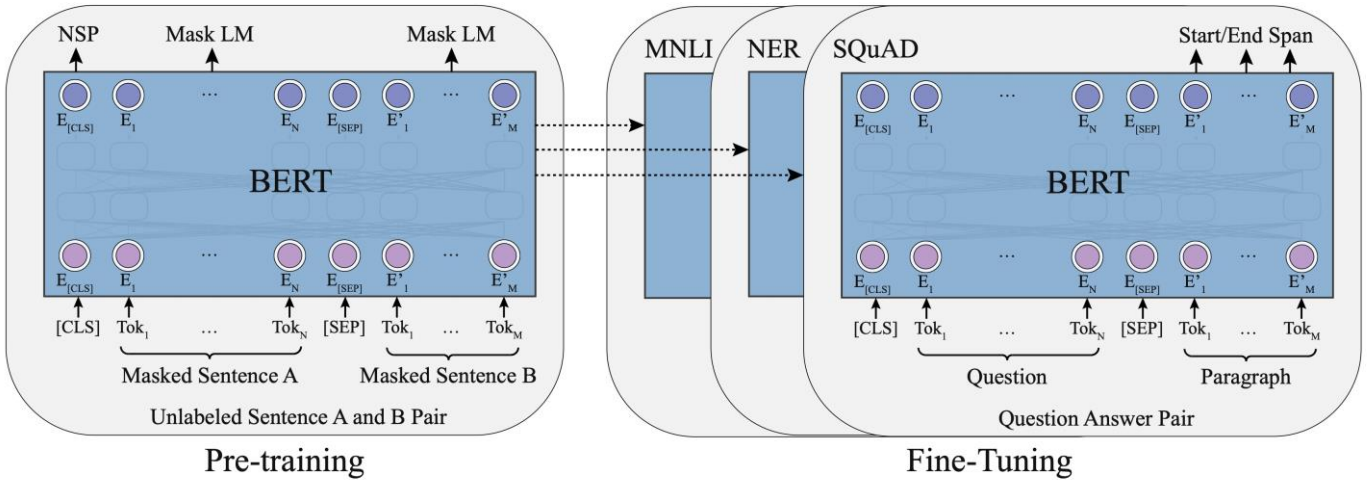
  - 核心思路：用大规模语料训练双向语言模型，得到能根据上下文语境变换而改变的词向量

ELMo: Deep Contextual Word Embeddings. NAACL 2018

Fine-tuning-based PLMs

- BERT
  - 采用了 Transformer 来对文本进行编码
  - 提出 Masked Language Model 进行预训练
  - 在多数常见 NLP 任务上效果显著





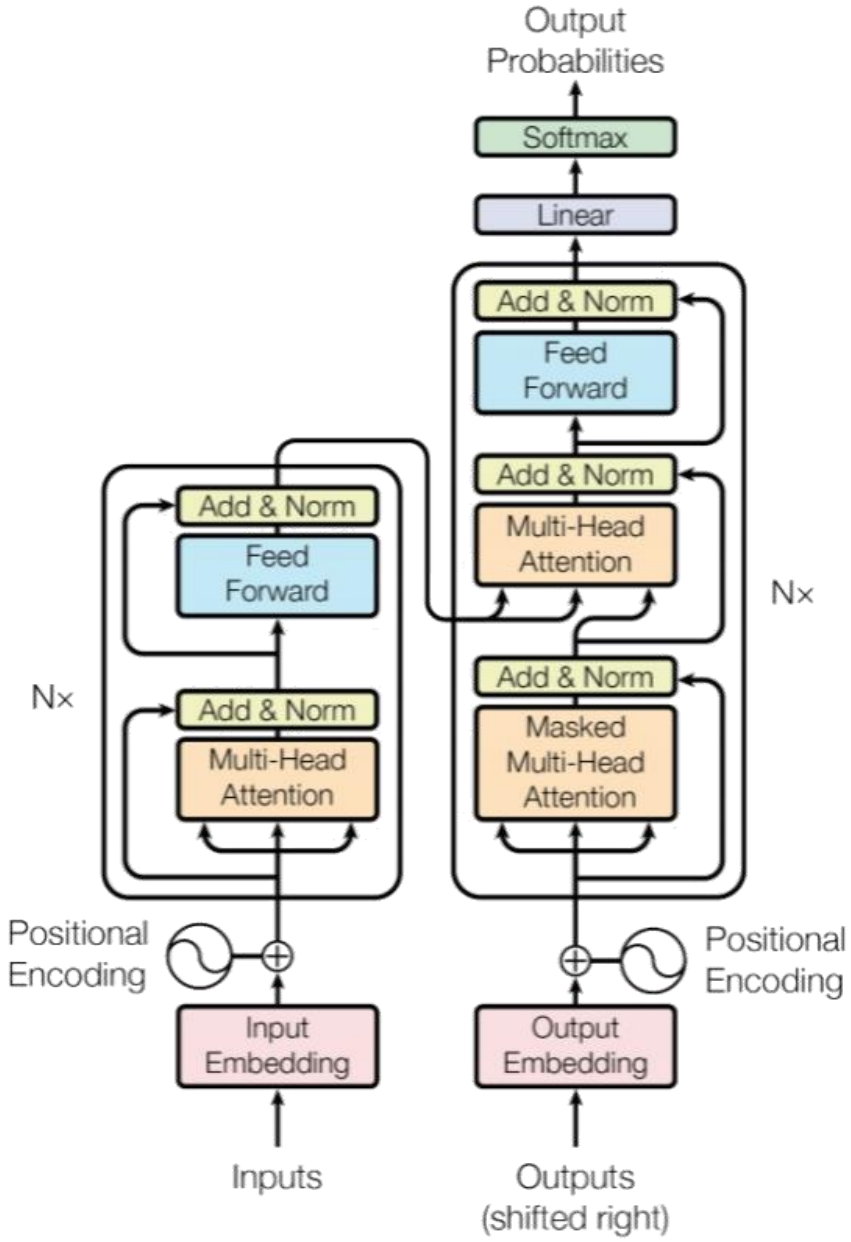
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019

Fine-tuning-based PLMs

优点：

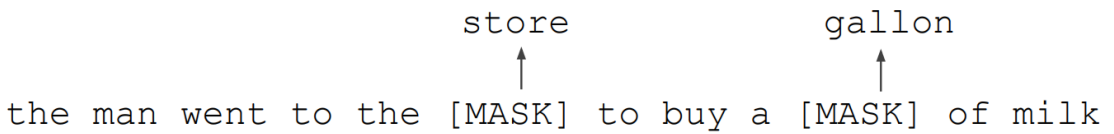
- Total computational complexity per layer
- Amount of computation that can be parallelized, as mesured by the minimum number of sequential operations required
- Path length between long-range dependencies in the network

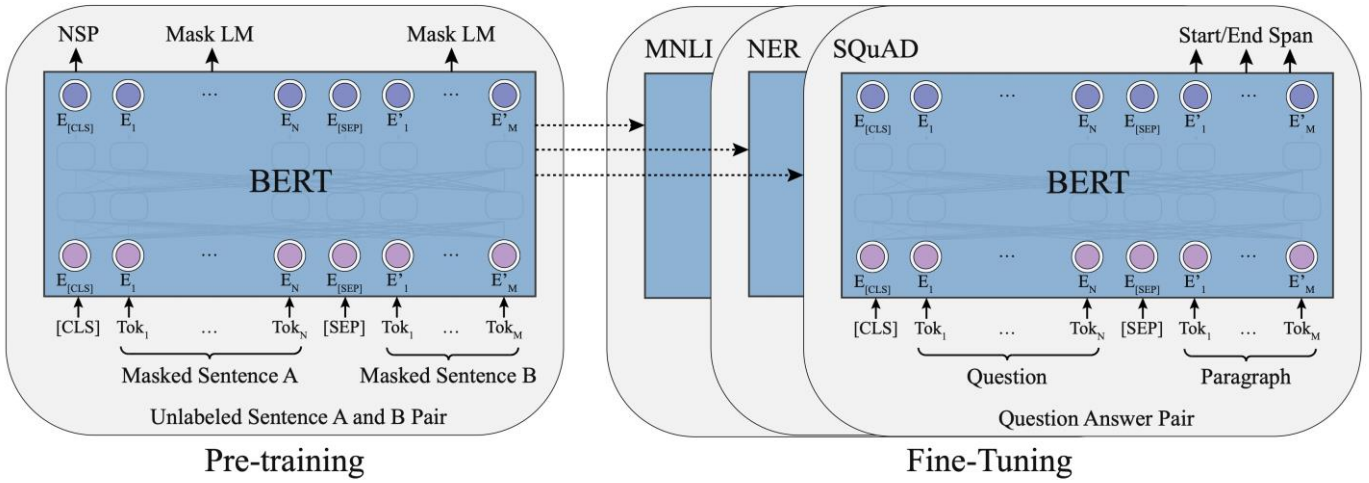
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$



Fine-tuning-based PLMs

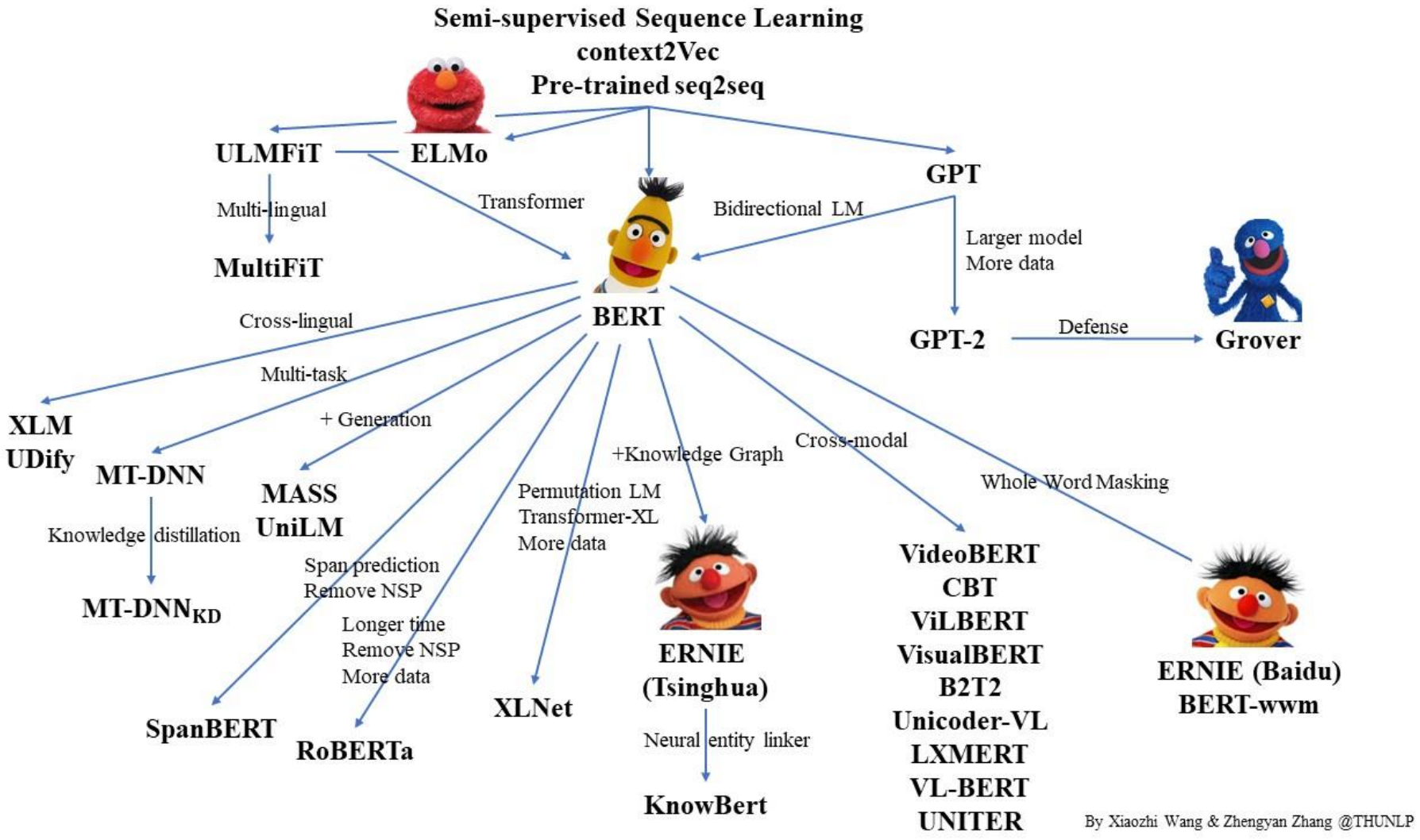
- BERT
  - 采用了 Transformer 来对文本进行编码
  - 提出 Masked Language Model 进行预训练
  - 在多数常见 NLP 任务上效果显著





BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019

Pre-trained Language Models (PLMs)



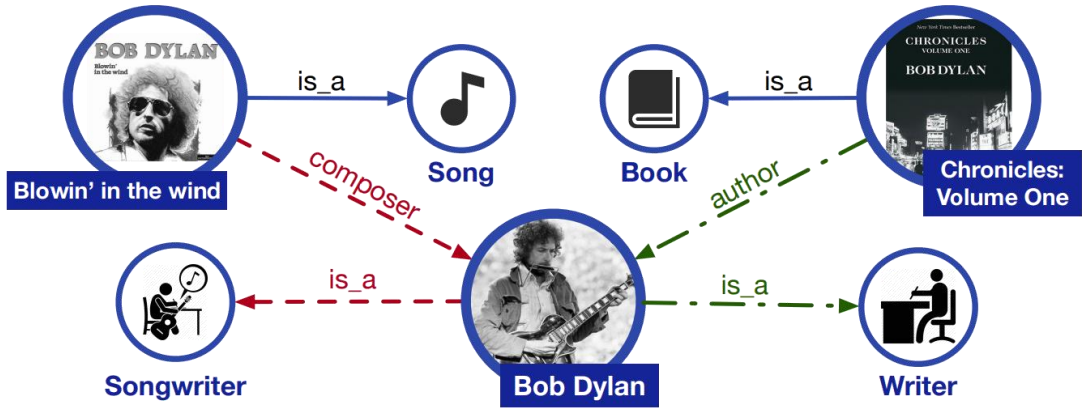
https://github.com/thunlp/PLMpapers

PLMs with Knowledge

- ERNIE (THU)
- KEPLER
- KnowBERT
- CoLAKE
- K-BERT
- WKLM
- K-Adapter

PLMs with Knowledge (Motivation)

- 现有预训练语言模型难以捕获**低频实体**信息
  - ELMo: Character CNN
  - BERT: sub-word
- 外部知识可以增强 PLMs
- 有助于一些**知识驱动**的下游任务
  - 关系分类 (Relation Classification)
  - 实体分类 (Entity Typing)

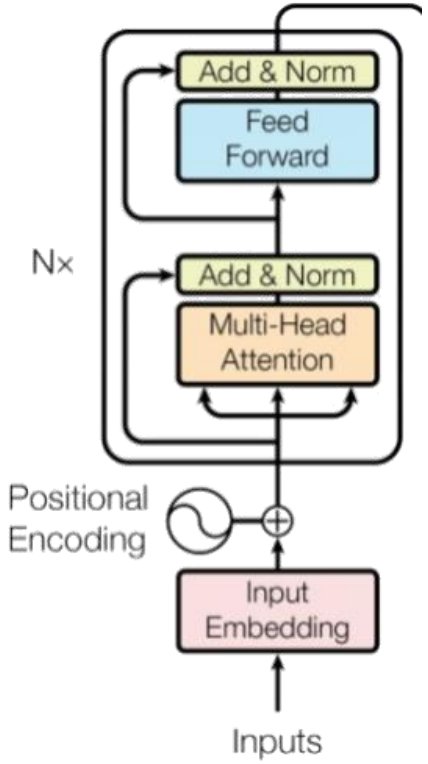


Bob Dylan wrote ***Blowin' in the Wind*** in 1962, and wrote ***Chronicles: Volume One*** in 2004.

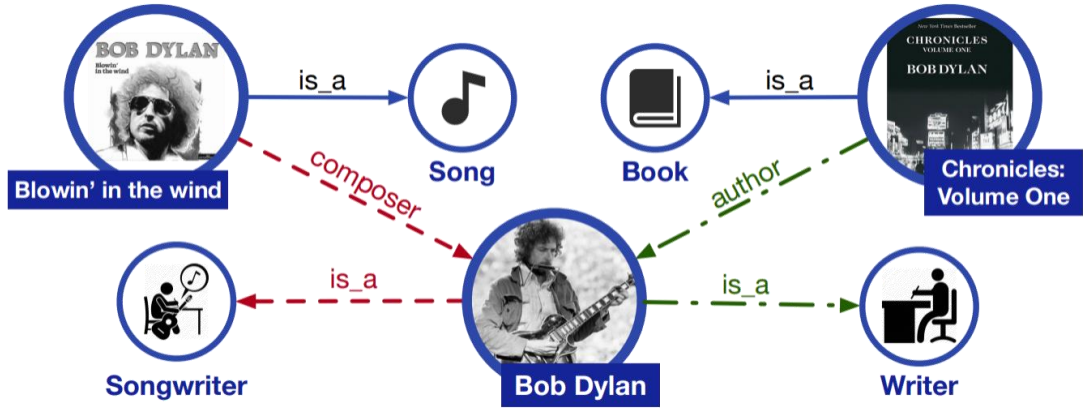
PLMs with Knowledge (Challenges)

- **结构化知识表示 (Structured Knowledge Encoding)**
  - 根据文本从知识图谱中检索相关知识
  - 将结构化信息表示为低维向量
- **异质信息融合 (Heterogeneous Information Fusion)**
  - 自然语言：词法、句法、语义
  - 知识图谱：实体、关系

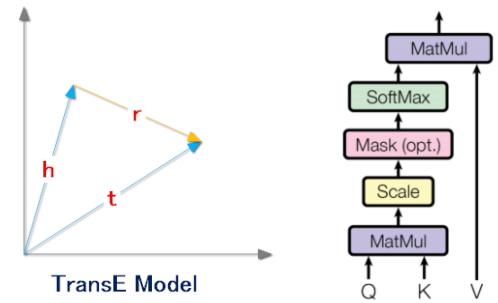
PLMs with Knowledge (Precondition)

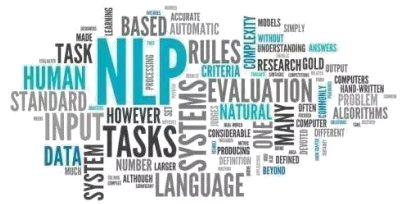


预训练语言模型（BERT、XLNet 等）



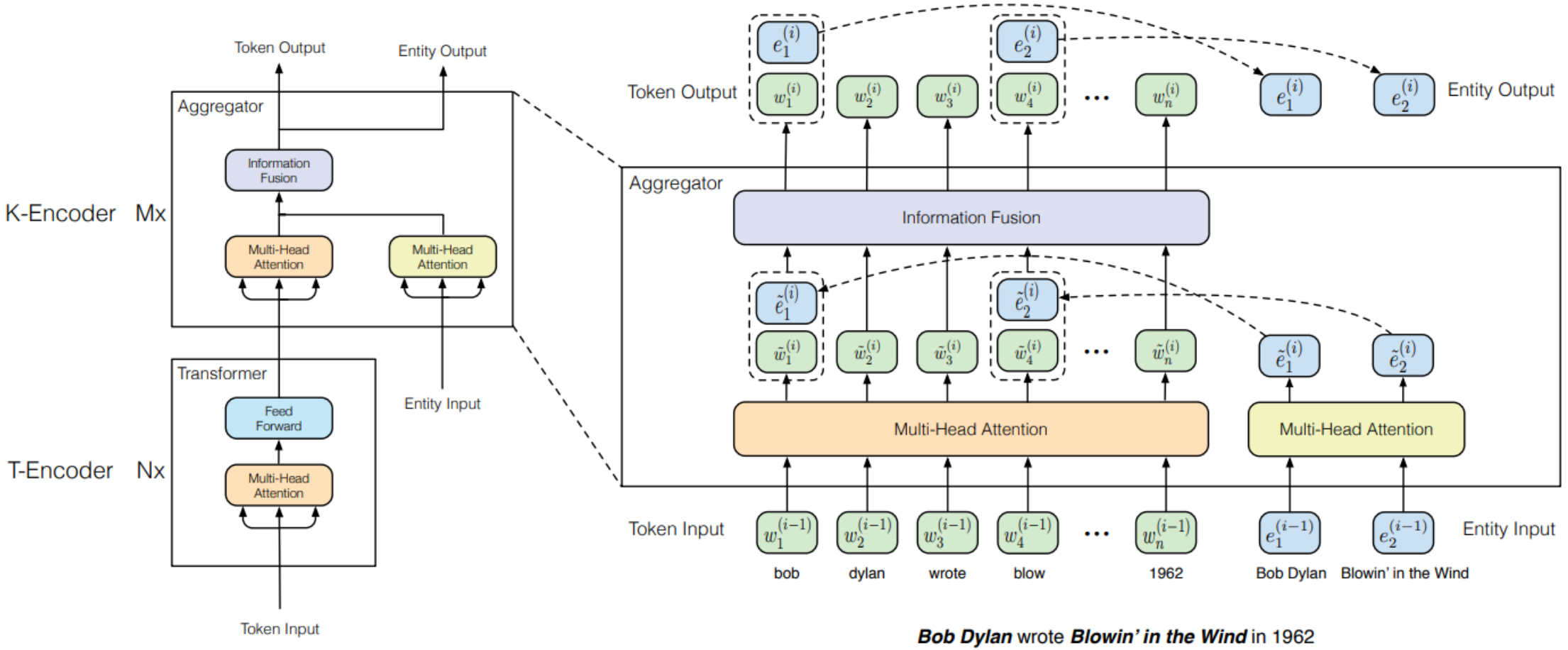
知识图谱（WordNet、Wikidata 等）





一系列技术

ERNIR (THU)



(a) Model Achitecture

(b) Aggregator

ERNIE: Enhanced Language Representation with Informative Entities. ACL 2019

ERNIR (THU)

Model	Acc.	Macro	Micro
NFGEC (Attentive)	54.53	74.76	71.58
NFGEC (LSTM)	55.60	75.15	71.73
BERT	52.04	75.16	71.63
ERNIE	**57.19**	**76.51**	**73.39**

Table 2: Results of various models on FIGER (%).

Model	P	R	F1
NFGEC (LSTM)	68.80	53.30	60.10
UFET	77.40	60.60	68.00
BERT	76.37	70.96	73.56
ERNIE	**78.42**	**72.90**	**75.56**

Table 3: Results of various models on Open Entity (%).

Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	**88.32**	69.97	66.08	**67.97**

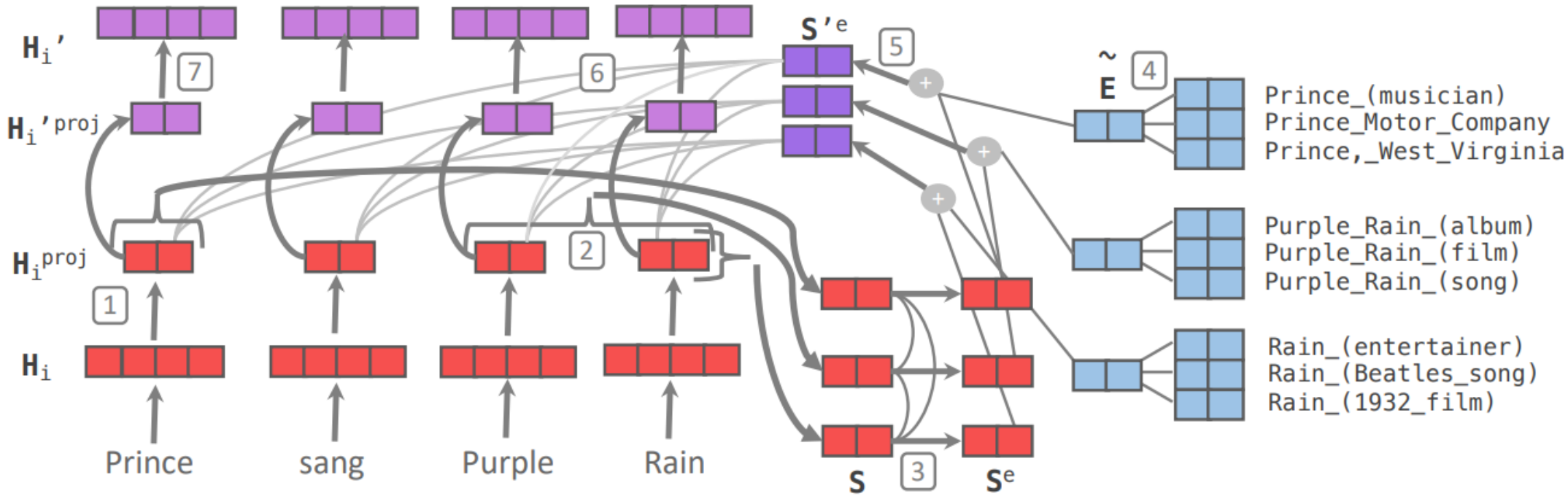
Table 5: Results of various models on FewRel and TACRED (%).

Model	MNLI-(m/mm) 392k	QQP 363k	QNLI 104k	SST-2 67k
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	-	93.5
ERNIE	84.0/83.2	71.2	91.3	93.5
Model	**CoLA 8.5k**	**STS-B 5.7k**	**MRPC 3.5k**	**RTE 2.5k**
$\text{BERT}_{\text{BASE}}$	52.1	85.8	88.9	66.4
ERNIE	52.3	83.2	88.2	68.8

Table 6: Results of BERT and ERNIE on different tasks of GLUE (%).

ERNIE: Enhanced Language Representation with Informative Entities. ACL 2019

KnowBERT



Knowledge Enhanced Contextual Word Representations. EMNLP 2019

KnowBERT

System	Accuracy
ELMo†	57.7
$\text{BERT}_{\text{BASE}}$ †	65.4
$\text{BERT}_{\text{LARGE}}$ †	65.5
$\text{BERT}_{\text{LARGE}}$ ††	69.5
KnowBert-W+W	**70.9**

Table 6: Test set results for the WiC dataset (v1.0).
†Pilehvar and Camacho-Collados (2019)
††Wang et al. (2019a)

System	P	R	F_1
UFET	68.8	53.3	60.1
$\text{BERT}_{\text{BASE}}$	76.4	71.0	73.6
ERNIE	78.4	72.9	75.6
KnowBert-W+W	78.6	73.7	**76.1**

Table 7: Test set results for entity typing using the nine general types from (Choi et al., 2018).

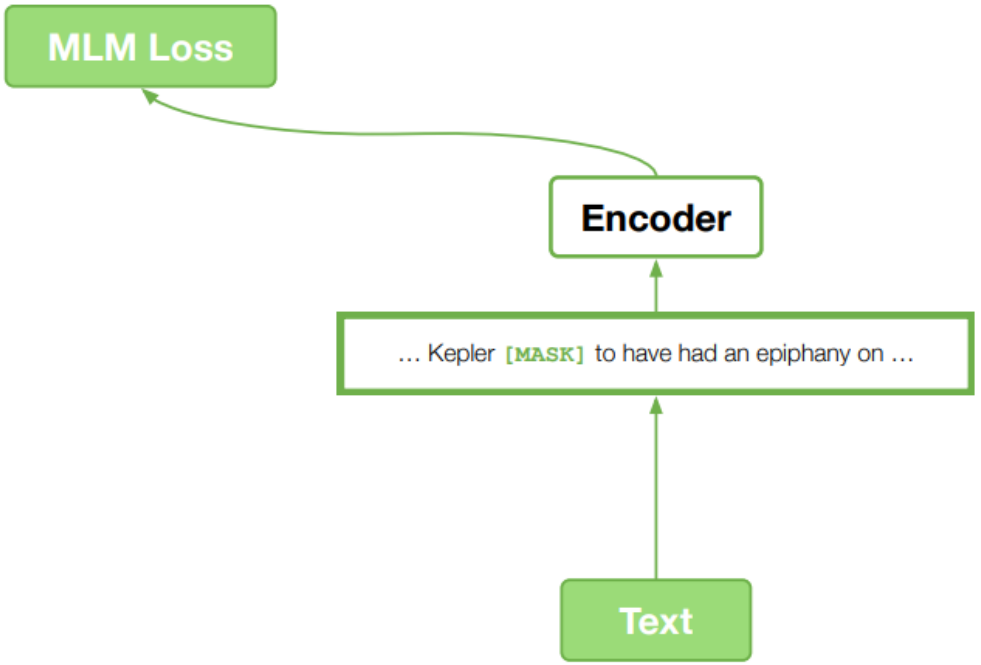
Knowledge Enhanced Contextual Word Representations. EMNLP 2019

A Different Direction

- 使用训练好的图谱表示作为输入存在一些问题
  - 图谱表示空间**难以**和语言表示空间融合
  - 需要实体链接工具，带来额外开销以及可能链接错误

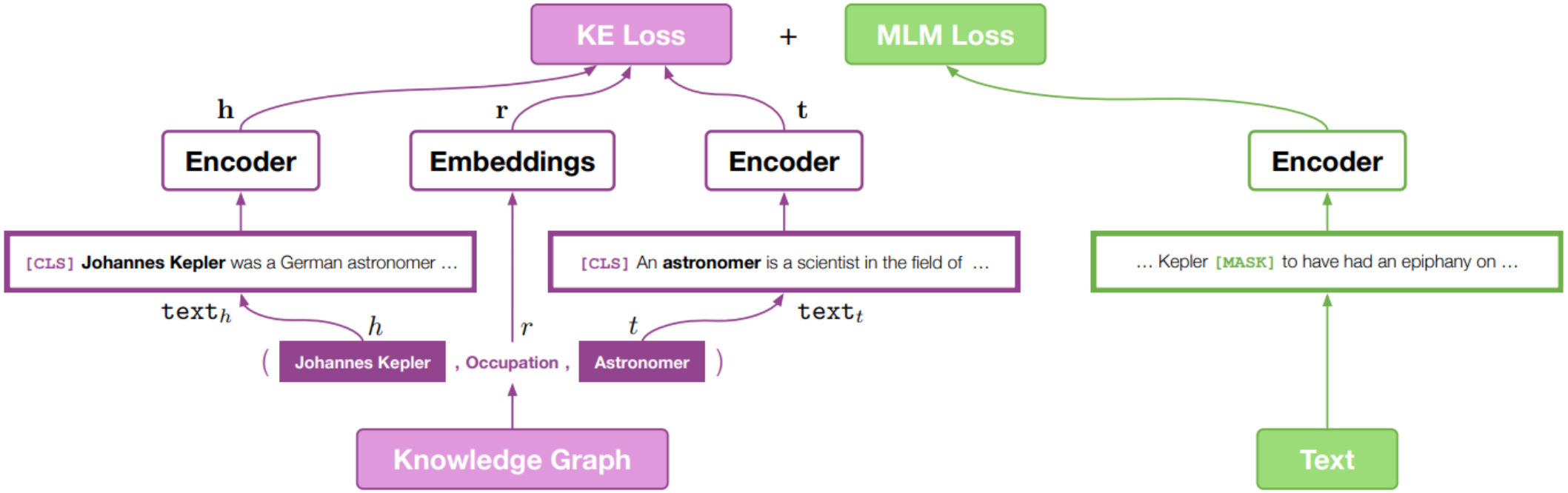
KEPLER

普通的 PLMs



KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. TACL

KEPLER



KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. TACL

KEPLER

Method	MR	MRR	HITS@1	HITS@3	HITS@10
TransE (Bordes et al., 2013)	109370	25.3	17.0	31.1	39.2
DistMult (Yang et al., 2015)	211030	25.3	20.8	27.8	33.4
ComplEx (Trouillon et al., 2016)	244540	28.1	22.8	31.0	37.3
SimplE (Kazemi and Poole, 2018)	115263	29.6	25.2	31.7	37.7
RotatE (Sun et al., 2019)	89459	29.0	23.4	32.2	39.0

Table 4: Performances of different KE models on Wikidata5M (% except MR).

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. TACL

KEPLER

Model	P	R	F-1
BERT	67.2	64.8	66.0
$\text{BERT}_{\text{LARGE}}$	-	-	70.1
MTB	69.7	67.9	68.8
MTB ($\text{BERT}_{\text{LARGE}}$)	-	-	71.5
$\text{ERNIE}_{\text{BERT}}$	70.0	66.1	68.0
$\text{KnowBert}_{\text{BERT}}$	**73.5**	64.1	68.5
RoBERTa	70.4	71.1	70.7
$\text{ERNIE}_{\text{RoBERTa}}$	**73.5**	68.0	70.7
$\text{KnowBert}_{\text{RoBERTa}}$	71.9	69.9	70.9
Our RoBERTa	70.8	69.6	70.2
KEPLER-Wiki	71.5	**72.5**	**72.0**
KEPLER-WordNet	71.4	71.3	71.3
KEPLER-W+W	71.1	72.0	71.5
KEPLER-Rel	71.3	70.9	71.1
KEPLER-Cond	72.1	70.7	71.4
KEPLER-OnlyDesc	72.3	69.1	70.7
KEPLER-KE	63.5	60.5	62.0

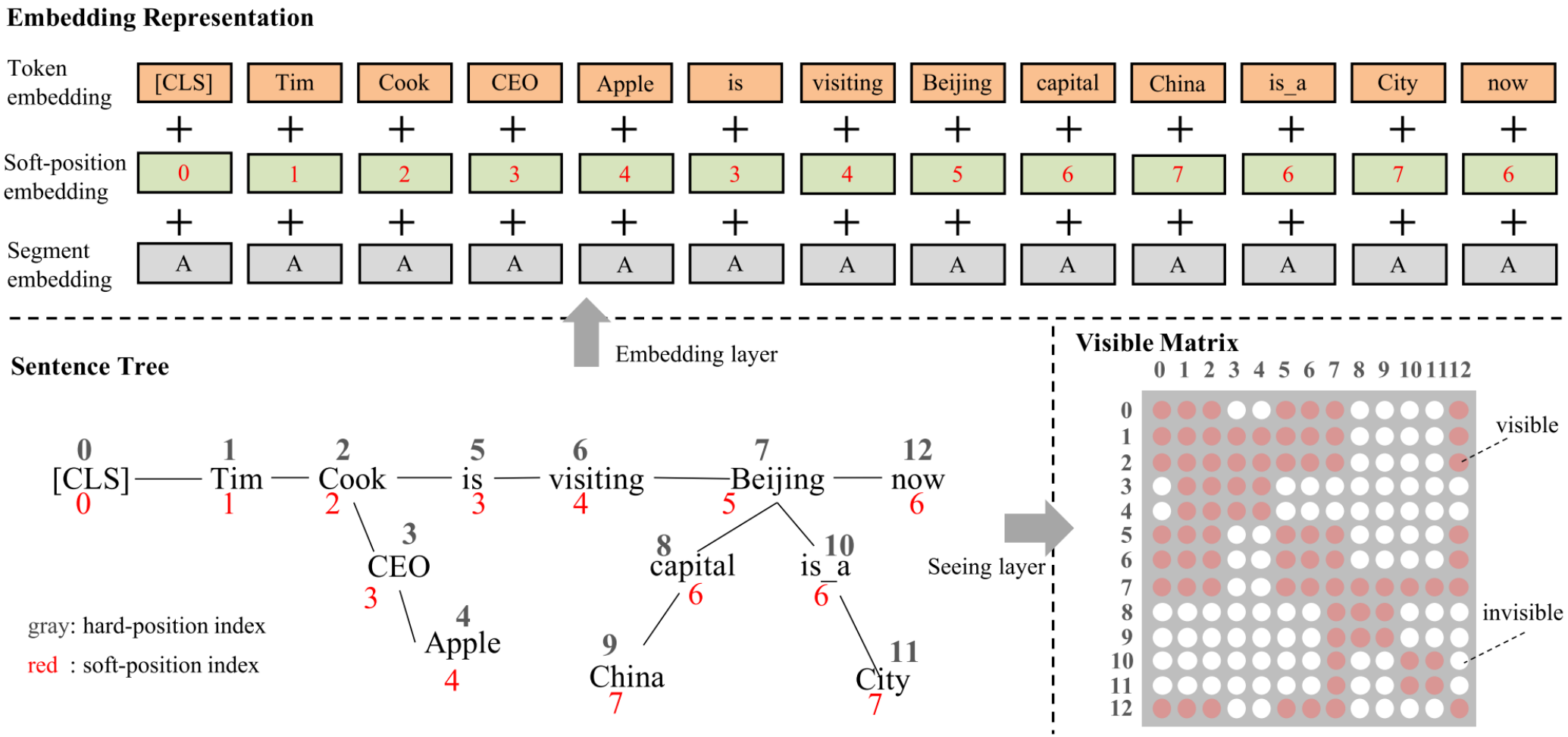
Table 5: Precision, recall and F-1 on TACRED (%). KnowBert results are different from the original paper since different task settings are used.

Model	P	R	F-1
UFET (Choi et al., 2018)	77.4	60.6	68.0
BERT	76.4	71.0	73.6
$\text{ERNIE}_{\text{BERT}}$	78.4	72.9	75.6
$\text{KnowBert}_{\text{BERT}}$	77.9	71.2	74.4
RoBERTa	77.4	73.6	75.4
$\text{ERNIE}_{\text{RoBERTa}}$	80.3	70.2	74.9
$\text{KnowBert}_{\text{RoBERTa}}$	78.7	72.7	75.6
Our RoBERTa	75.1	73.4	74.3
KEPLER-Wiki	77.8	74.6	**76.2**

Table 7: Entity typing results on OpenEntity (%).

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. TACL

K-BERT



K-BERT: Enabling Language Representation with Knowledge Graph. AAAI 2020

K-BERT

Table 1: Results of various models on sentence classification tasks on open-domain tasks (*Acc. %*)

Models\Datasets	**Book_review**		**Chnsenticorp**		**Shopping**		**Weibo**		**XNLI**		**LCQMC**	
	Dev	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
Pre-trainied on WikiZh by Google.												
Google BERT	88.3	**87.5**	93.3	94.3	96.7	96.3	98.2	98.3	76.0	75.4	88.4	86.2
K-BERT (HowNet)	88.6	87.2	**94.6**	**95.6**	**97.1**	**97.0**	98.3	98.3	**76.8**	**76.1**	**88.9**	86.9
K-BERT (CN-DBpedia)	**88.6**	87.3	93.9	95.3	96.6	96.5	98.3	98.3	76.5	76.0	88.6	**87.0**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	**88.6**	87.9	94.8	95.7	96.9	**97.1**	98.2	98.2	77.0	76.3	89.0	86.7
K-BERT (HowNet)	88.5	87.4	**95.4**	95.6	96.9	96.9	98.3	**98.4**	**77.2**	**77.0**	**89.2**	**87.1**
K-BERT (CN-DBpedia)	88.8	87.9	95.0	**95.8**	**97.1**	97.0	98.3	98.3	76.2	75.9	89.0	86.9

K-BERT: Enabling Language Representation with Knowledge Graph. AAAI 2020

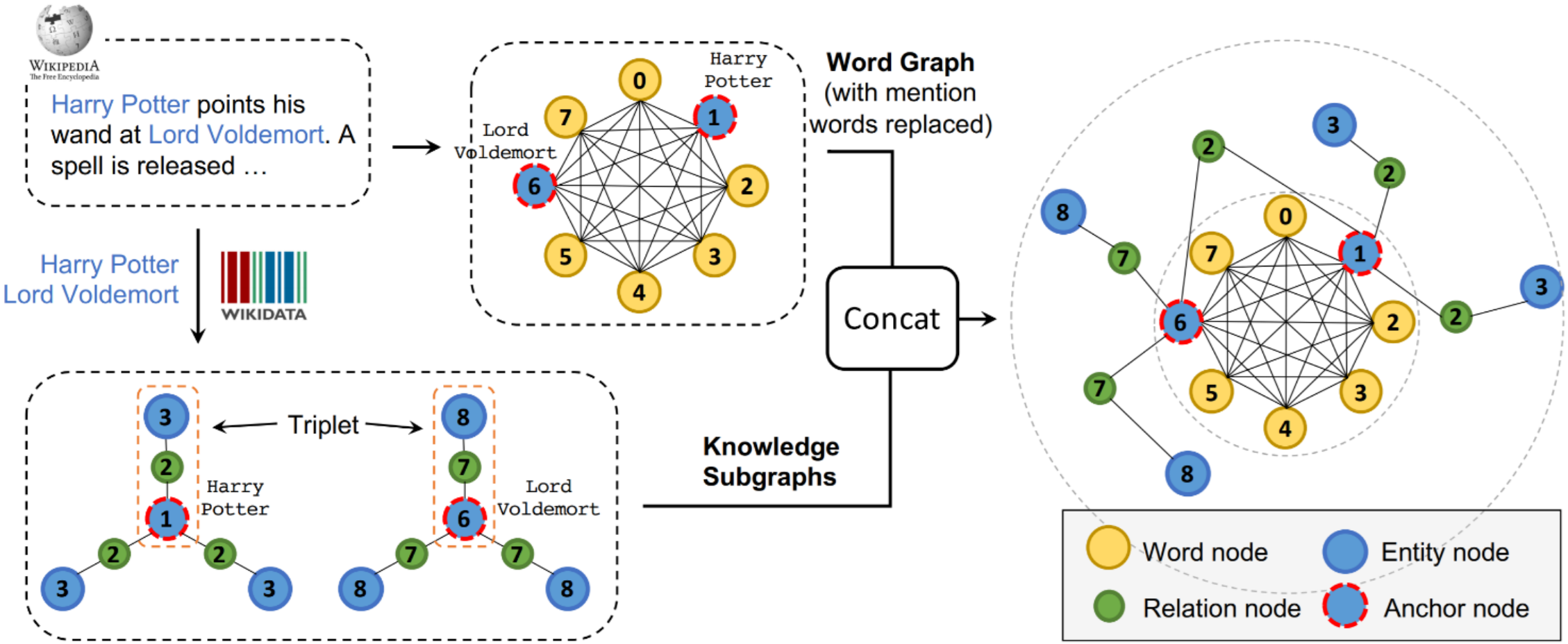
K-BERT

Table 3: Results of various models on specific-domain tasks (%).

Models\Datasets	Finance_Q&A			Law_Q&A			Finance_NER			Medicine_NER		
	$P.$	$R.$	$F1$	$P.$	$R.$	$F1$	$P.$	$R.$	$F1$	$P.$	$R.$	$F1$
Pre-trained on WikiZh by Google.												
Google BERT	81.9	86.0	83.9	83.1	90.1	86.4	84.8	87.4	86.1	91.9	93.1	92.5
K-BERT (HowNet)	83.3	84.4	83.9	83.7	91.2	87.3	86.3	89.0	**87.6**	93.2	93.3	93.3
K-BERT (CN-DBpedia)	81.5	88.6	**84.9**	82.1	93.8	**87.5**	86.1	88.7	87.4	93.9	93.8	93.8
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.0	94.4	**94.2**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	82.1	86.5	84.2	83.2	91.7	87.2	84.9	87.4	86.1	91.8	93.5	92.7
K-BERT (HowNet)	82.8	85.8	84.3	83.0	92.4	87.5	86.3	88.5	87.3	93.5	93.8	93.7
K-BERT (CN-DBpedia)	81.9	87.1	**84.4**	83.1	92.6	**87.6**	86.3	88.6	**87.4**	93.9	94.3	94.1
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.1	94.3	**94.2**

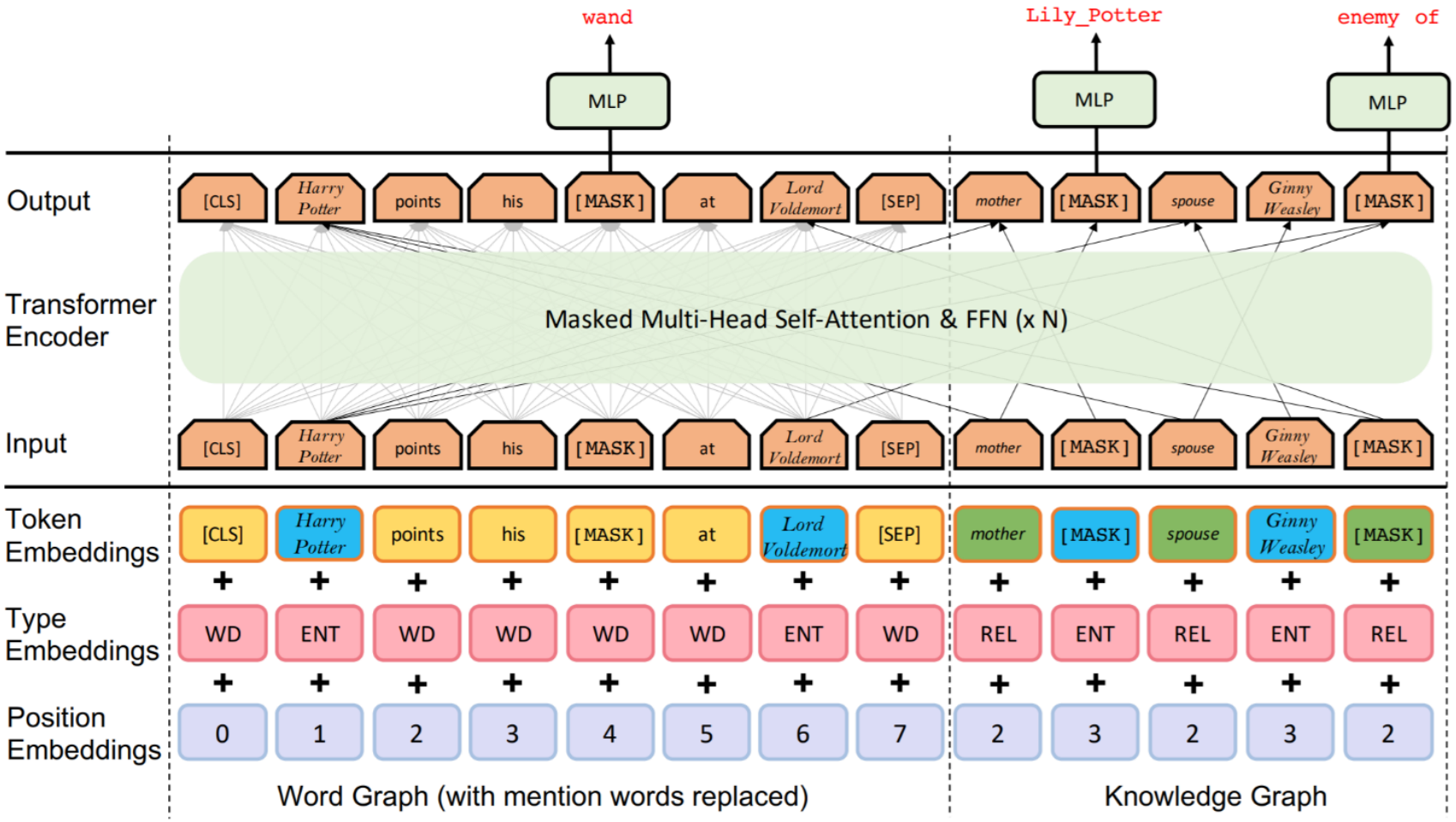
K-BERT: Enabling Language Representation with Knowledge Graph. AAAI 2020

CoLAKE



CoLAKE: Contextualized Language and Knowledge Embedding. COLING 2020

CoLAKE



CoLAKE: Contextualized Language and Knowledge Embedding. COLING 2020

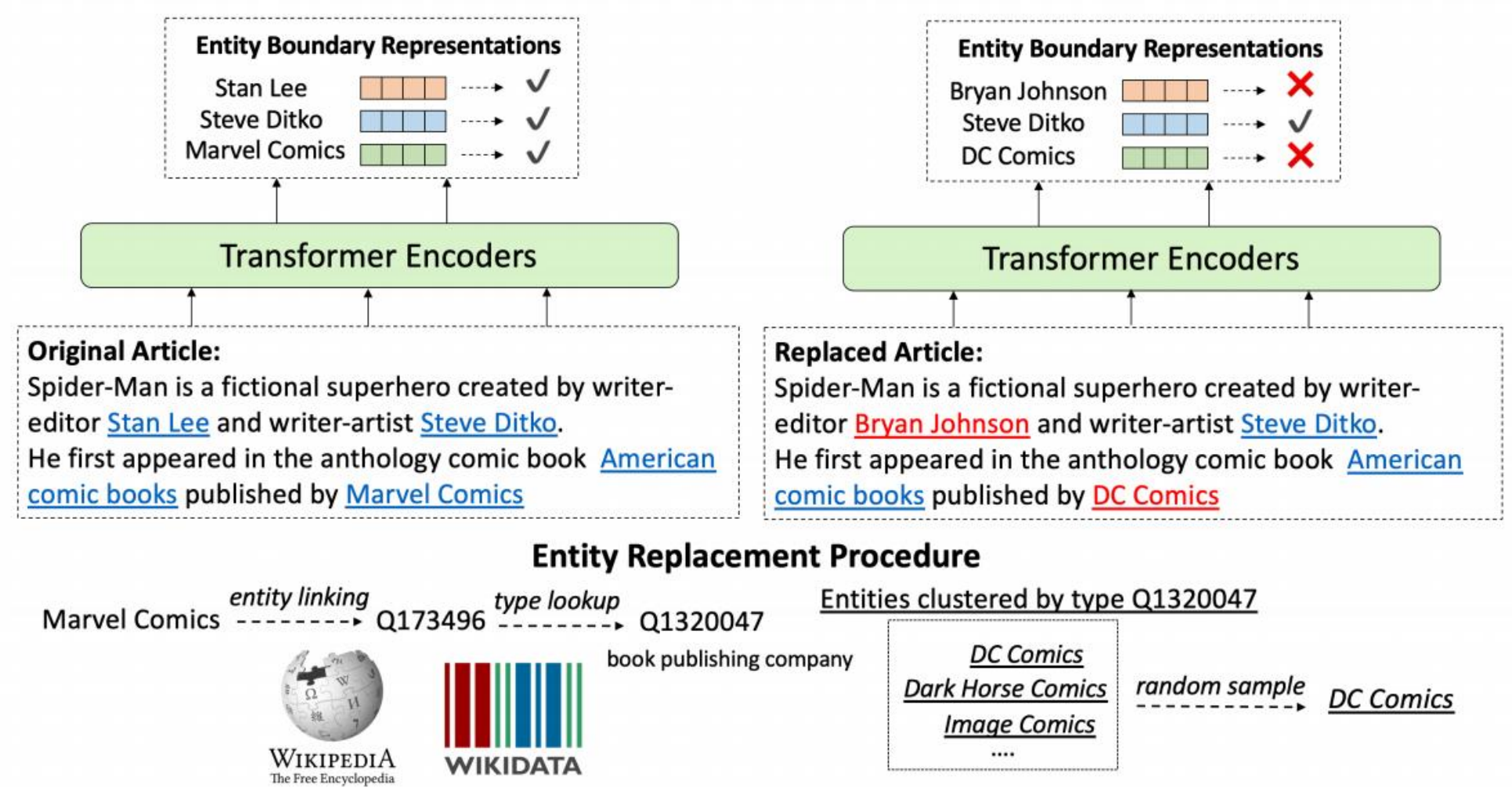
CoLAKE

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**

Table 2: Experimental results on Open Entity and FewRel.

CoLAKE: Contextualized Language and Knowledge Embedding. COLING 2020

WKLM (Pretrained Encyclopedia)



Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. ICLR 2020

WKLM (Pretrained Encyclopedia)

Table 4: Open-domain QA Results.

Model	WebQuestions		TriviaQA		Quasar-T		SearchQA	
	EM	F1	EM	F1	EM	F1	EM	F1
DrQA (Chen et al., 2017)	20.7	-	-	-	-	-	-	-
R^3 (Wang et al., 2018a)	-	-	50.6	57.3	42.3	49.6	57.0	63.2
DSQA (Lin et al., 2018)	18.5	25.6	48.7	56.3	42.2	49.3	49.0	55.3
Evidence Agg. (Wang et al., 2018b)	-	-	50.6	57.3	42.3	49.6	57.0	63.2
BERTserini (Yang et al., 2019a)	-	-	51.0	56.3	-	-	-	-
BERTserini+DS (Yang et al., 2019b)	-	-	54.4	60.2	-	-	-	-
ORQA (Lee et al., 2019)	**36.4**	-	45.0	-	-	-	-	-
Our BERT	29.2	35.5	48.7	53.2	40.4	46.1	57.1	61.9
Our BERT + Ranking score	32.2	38.9	52.1	56.5	43.2	49.2	60.6	65.9
WKLM	30.8	37.9	52.2	56.7	43.7	49.9	58.7	63.3
WKLM + Ranking score	34.6	41.8	**58.1**	**63.1**	**45.8**	**52.2**	**61.7**	**66.7**

Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. ICLR 2020

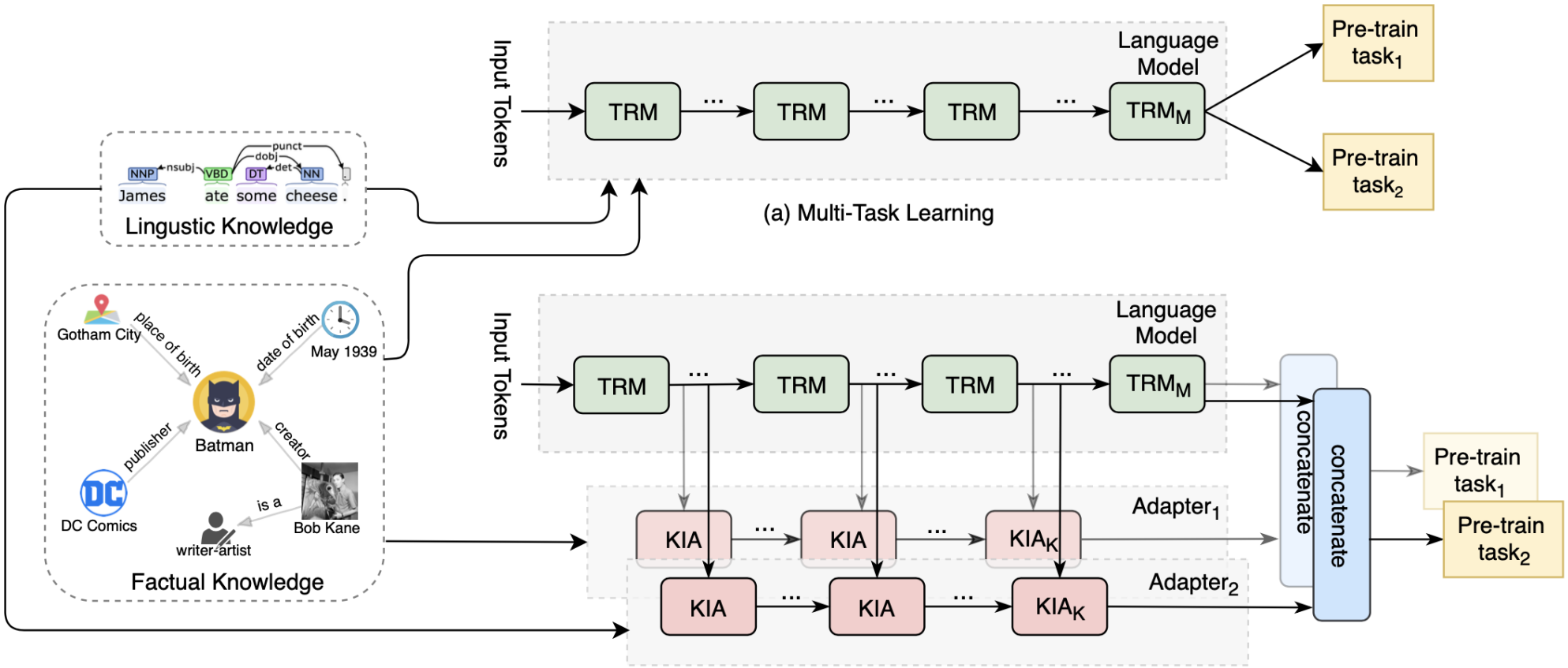
WKLM (Pretrained Encyclopedia)

Table 5: Fine-grained Entity Typing Results on the FIGER dataset.

Model	Acc	Ma-F1	Mi-F1
LSTM + Hand-crafted (Inui et al., 2017)	57.02	76.98	73.94
Attentive + Hand-crafted (Inui et al., 2017)	59.68	78.97	75.36
BERT baseline (Zhang et al., 2019)	52.04	75.16	71.63
ERNIE (Zhang et al., 2019)	57.19	75.61	73.39
Our BERT	54.53	79.57	74.74
WKLM	**60.21**	**81.99**	**77.00**

Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. ICLR 2020

K-Adapter



以往的任务：

- 无法 continual learning
- 不同类型知识注入之间是耦合的

K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. arXiv preprint arXiv:2002.01808, 2020

K-Adapter

Model	SearchQA		Quasar-T		CosmosQA
	EM	F_1	EM	F_1	Accuracy
BiDAF (Seo et al., 2016)	28.60	34.60	25.90	28.50	-
AQA (Buck et al., 2018)	40.50	47.40	-	-	-
Rˆ3 (Wang et al., 2017a)	49.00	55.30	35.30	41.70	-
DSQA (Lin et al., 2018)	49.00	55.30	42.30	49.30	-
Evidence Agg. (Wang et al., 2018)	57.00	63.20	42.30	49.60	-
BERT (Xiong et al., 2020)	57.10	61.90	40.40	46.10	-
WKLM (Xiong et al., 2020)	58.70	63.30	43.70	49.90	-
WKLM + Ranking (Xiong et al., 2020)	61.70	66.70	45.80	52.20	-
BERT-FT$_{RACE+SWAG}$ (Huang et al., 2019)	-	-	-	-	68.70
RoBERTa	59.01	65.62	40.83	48.84	80.59
RoBERTa + multitask	59.92	66.67	44.62	51.17	81.19
K-ADAPTER (F)	61.85	67.17	46.20	52.86	80.93
K-ADAPTER (L)	61.15	66.82	45.66	52.39	80.76
K-ADAPTER (F+L)	**61.96**	**67.31**	**46.32**	**53.00**	**81.83**

Table 3: Results on question answering datasets including: CosmosQA, SearchQA and Quasar-T.

K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. arXiv preprint arXiv:2002.01808, 2020

Conclusion

- 将图谱表示向量作为特征输入（ERNIE、KnowBERT）
- 设计新的预训练任务（KEPLER、K-BERT、CoLAKE、WKLM）
- 增加额外的模块（K-Adapter）

THANKS