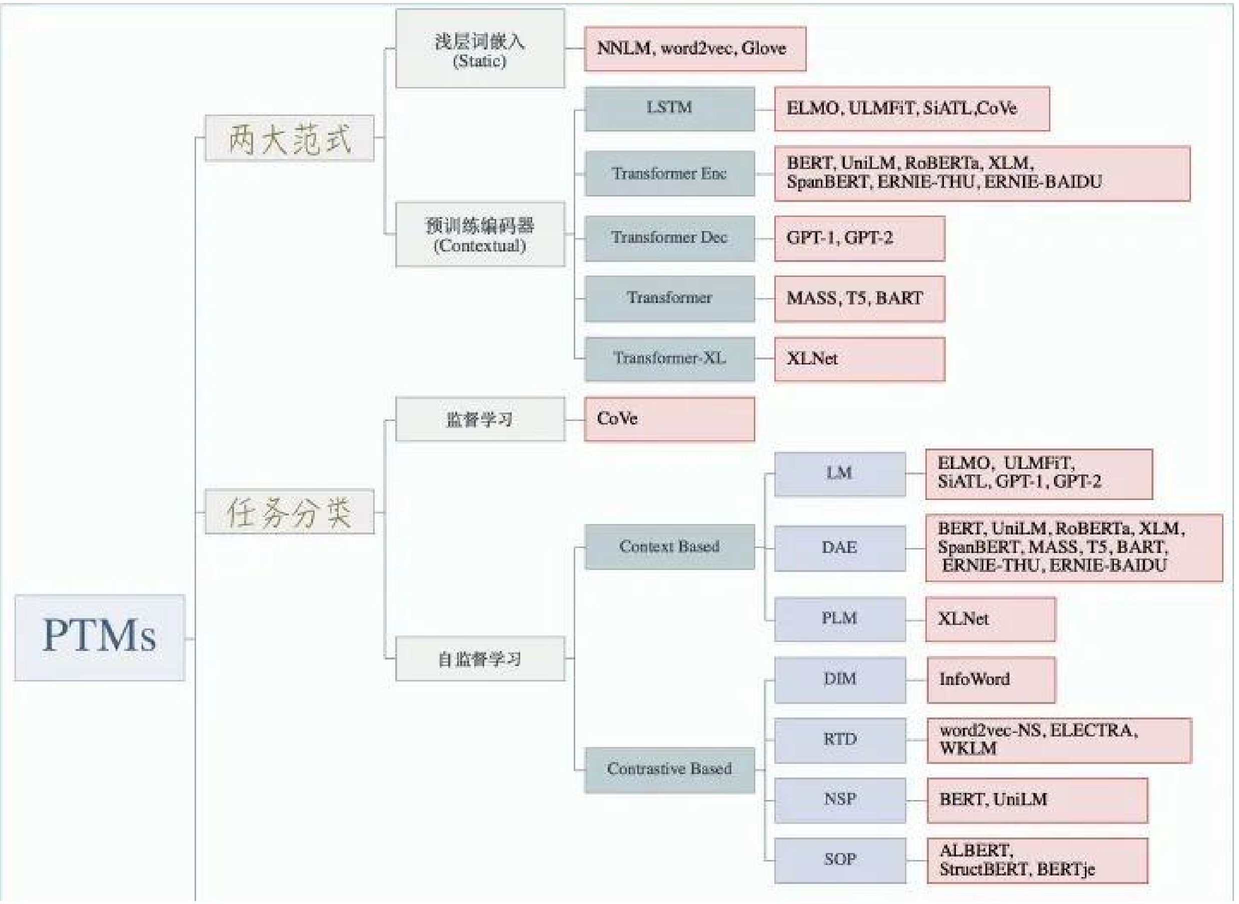
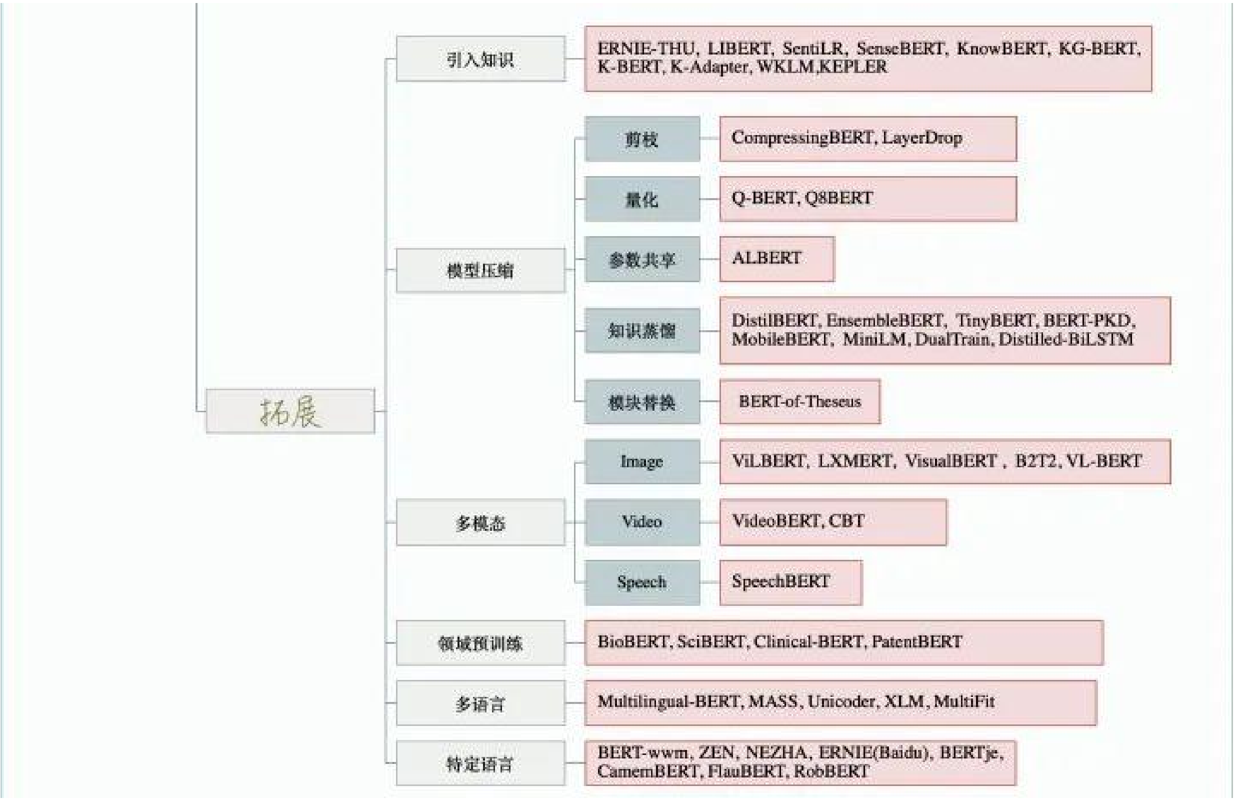
PTMs



- PTMs
  - 两大范式
    - 浅层词嵌入 (Static)
      - NNLM, word2vec, Glove
    - 预训练编码器 (Contextual)
      - LSTM
        - ELMO, ULMFiT, SiATL,CoVe
      - Transformer Enc
        - BERT, UniLM, RoBERTa, XLM, SpanBERT, ERNIE-THU, ERNIE-BAIDU
      - Transformer Dec
        - GPT-1, GPT-2
      - Transformer
        - MASS, T5, BART
      - Transformer-XL
        - XLNet
  - 任务分类
    - 监督学习
      - CoVe
    - 自监督学习
      - Context Based
        - LM
          - ELMO, ULMFiT, SiATL, GPT-1, GPT-2
        - DAE
          - BERT, UniLM, RoBERTa, XLM, SpanBERT, MASS, T5, BART, ERNIE-THU, ERNIE-BAIDU
        - PLM
          - XLNet
      - Contrastive Based
        - DIM
          - InfoWord
        - RTD
          - word2vec-NS, ELECTRA, WKLM
        - NSP
          - BERT, UniLM
        - SOP
          - ALBERT, StructBERT, BERTje



- 拓展
  - 引入知识
    - ERNIE-THU, LIBERT, SentiLR, SenseBERT, KnowBERT, KG-BERT, K-BERT, K-Adapter, WKLM,KEPLER
  - 模型压缩
    - 剪枝
      - CompressingBERT, LayerDrop
    - 量化
      - Q-BERT, Q8BERT
    - 参数共享
      - ALBERT
    - 知识蒸馏
      - DistilBERT, EnsembleBERT, TinyBERT, BERT-PKD, MobileBERT, MiniLM, DualTrain, Distilled-BiLSTM
    - 模块替换
      - BERT-of-Theseus
  - 多模态
    - Image
      - ViLBERT, LXMERT, VisualBERT , B2T2, VL-BERT
    - Video
      - VideoBERT, CBT
    - Speech
      - SpeechBERT
  - 领域预训练
    - BioBERT, SciBERT, Clinical-BERT, PatentBERT
  - 多语言
    - Multilingual-BERT, MASS, Unicoder, XLM, MultiFit
  - 特定语言
    - BERT-wwm, ZEN, NEZHA, ERNIE(Baidu), BERTje, CamemBERT, FlauBERT, RobBERT

——2020/11/26 朱静丹

清源 CPM(Chinese Pretrained Models)

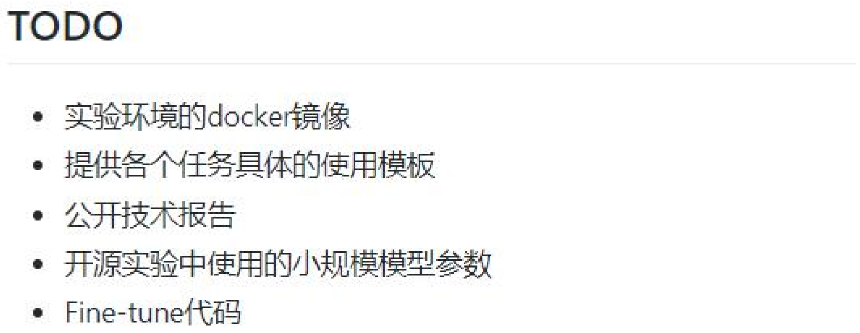
——北京智源人工智能研究院、清华大学研究团队

- 第一版 CPM 中文语言模型与 GPT-3 等预训练模型类似，通过少次、单次学习甚至零次学习，能完成不同自然语言处理任务，具备一定的常识和认知的泛化能力。

- CPM-LM 的参数规模为 26 亿，预训练中文数据规模100 GB，使用了 64 块 V100 GPU 训练时间约为 3 周。CPM-KG 的参数规模分别为217亿，预训练结构化知识图谱为 WikiData 全量数据，包含近 1300 个关系、8500万实体、4.8 亿个事实三元组，使用了 8 块 V100 GPU 训练时间约为 2 周

- https://github.com/TsinghuaAI/CPM-Generate
- https://cpm.baai.ac.cn/

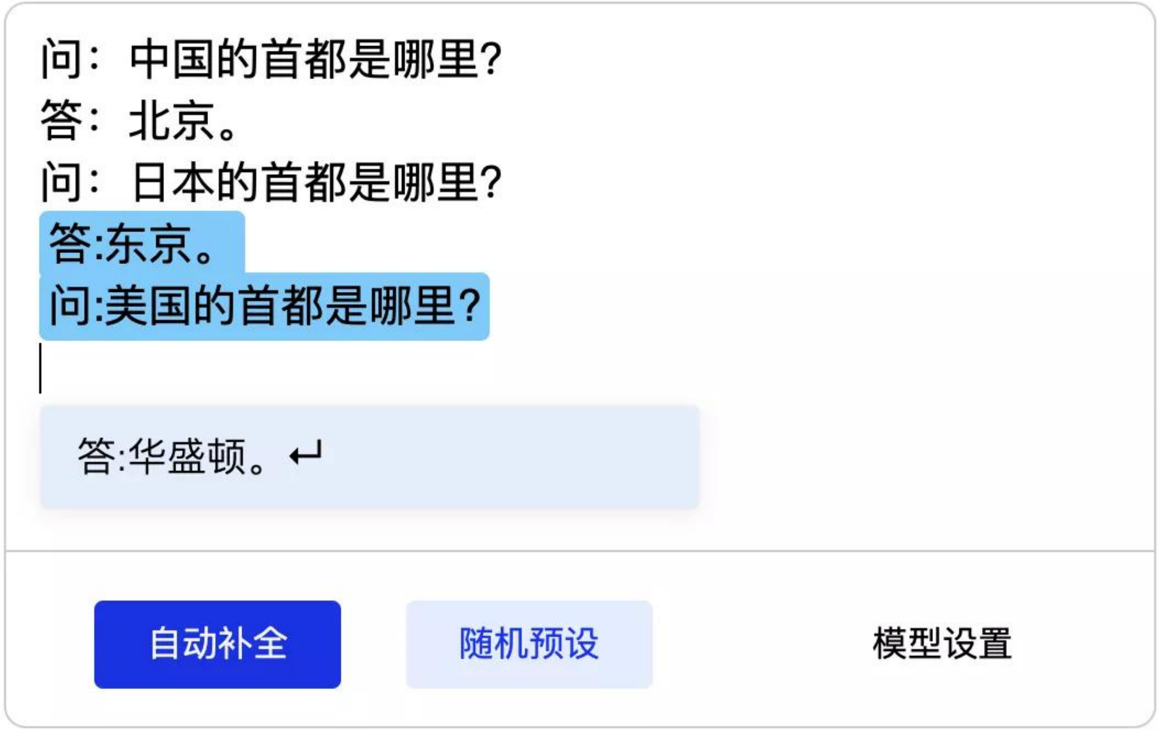
- https://github.com/qhduan/CPM-LM-TF2



TODO

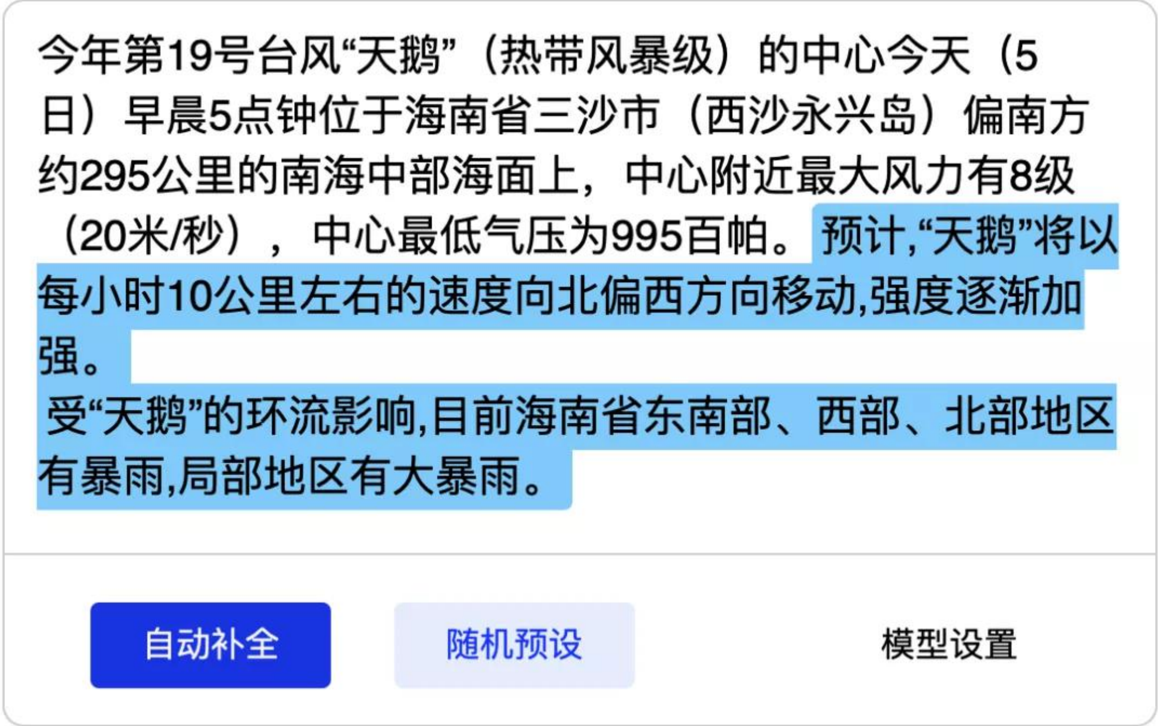
- 实验环境的docker镜像
- 提供各个任务具体的使用模板
- 公开技术报告
- 开源实验中使用的小规模模型参数
- Fine-tune代码

故事生成



问：中国的首都是哪里?
答：北京。
问：日本的首都是哪里?
答:东京。
问:美国的首都是哪里?

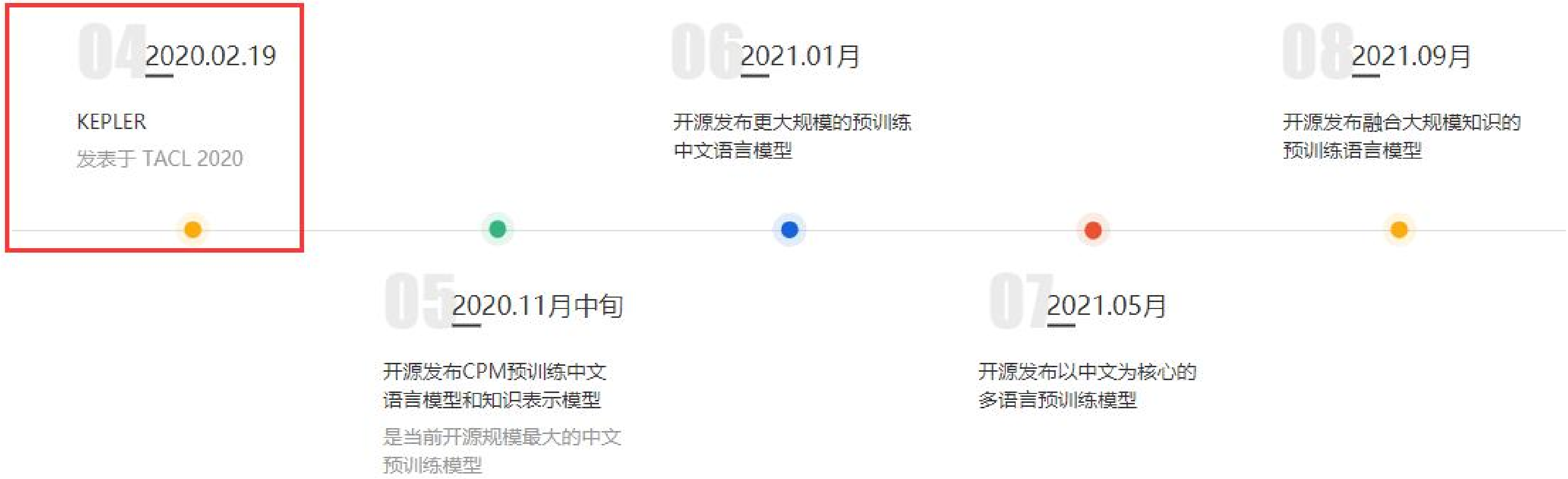
答:华盛顿。↵



今年第19号台风“天鹅”（热带风暴级）的中心今天（5日）早晨5点钟位于海南省三沙市（西沙永兴岛）偏南方约295公里的南海中部海面上，中心附近最大风力有8级（20米/秒），中心最低气压为995百帕。预计,“天鹅”将以每小时10公里左右的速度向北偏西方向移动,强度逐渐加强。

受“天鹅”的环流影响,目前海南省东南部、西部、北部地区有暴雨,局部地区有大暴雨。

历程规划



04 2020.02.19

KEPLER

发表于 TACL 2020

05 2020.11月中旬

开源发布CPM预训练中文语言模型和知识表示模型

是当前开源规模最大的中文预训练模型

06 2021.01月

开源发布更大规模的预训练中文语言模型

07 2021.05月

开源发布以中文为核心的多语言预训练模型

08 2021.09月

开源发布融合大规模知识的预训练语言模型

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

Xiaozhi Wang[1], Tianyu Gao[3], Zhaocheng Zhu[4,5], Zhengyan Zhang[1],
Zhiyuan Liu[1,2*], Juanzi Li[1,2], Jian Tang[4,6,7*]
[1]Department of CST, BNRist; [2]KIRC, Institute for AI, Tsinghua University, Beijing, China
{wangxz20,zy-z19}@mails.tsinghua.edu.cn
{liuzy,lijuanzi}@tsinghua.edu.cn
[3]Department of Computer Science, Princeton University, Princeton, USA
tianyug@princeton.edu
[4]Mila - Québec AI Institute; [5]Univesité de Montréal; [6]HEC, Montréal, Canada
zhaocheng.zhu@umontreal.ca, jian.tang@hec.ca
[7]CIFAR AI Research Chair

The contributions of this paper:

- This paper proposed a unified model for **K**nowledge **E**mbedding and **P**re-trained **L**anguag**E** **R**epresentation (KEPLER), which can not only better in tegrate factual knowledge into PLMs but also produce effective text-enhanced KE with the strong PLMs.

- They construct Wikidata5M[1], a large-scale KG dataset with aligned entity descriptions, and benchmark state-of-the-art KE methods on it.

- https://github.com/THU-KEG/KEPLER (None)
- https://deepgraphlearning.github.io/project/wikidata5m (Dataset)

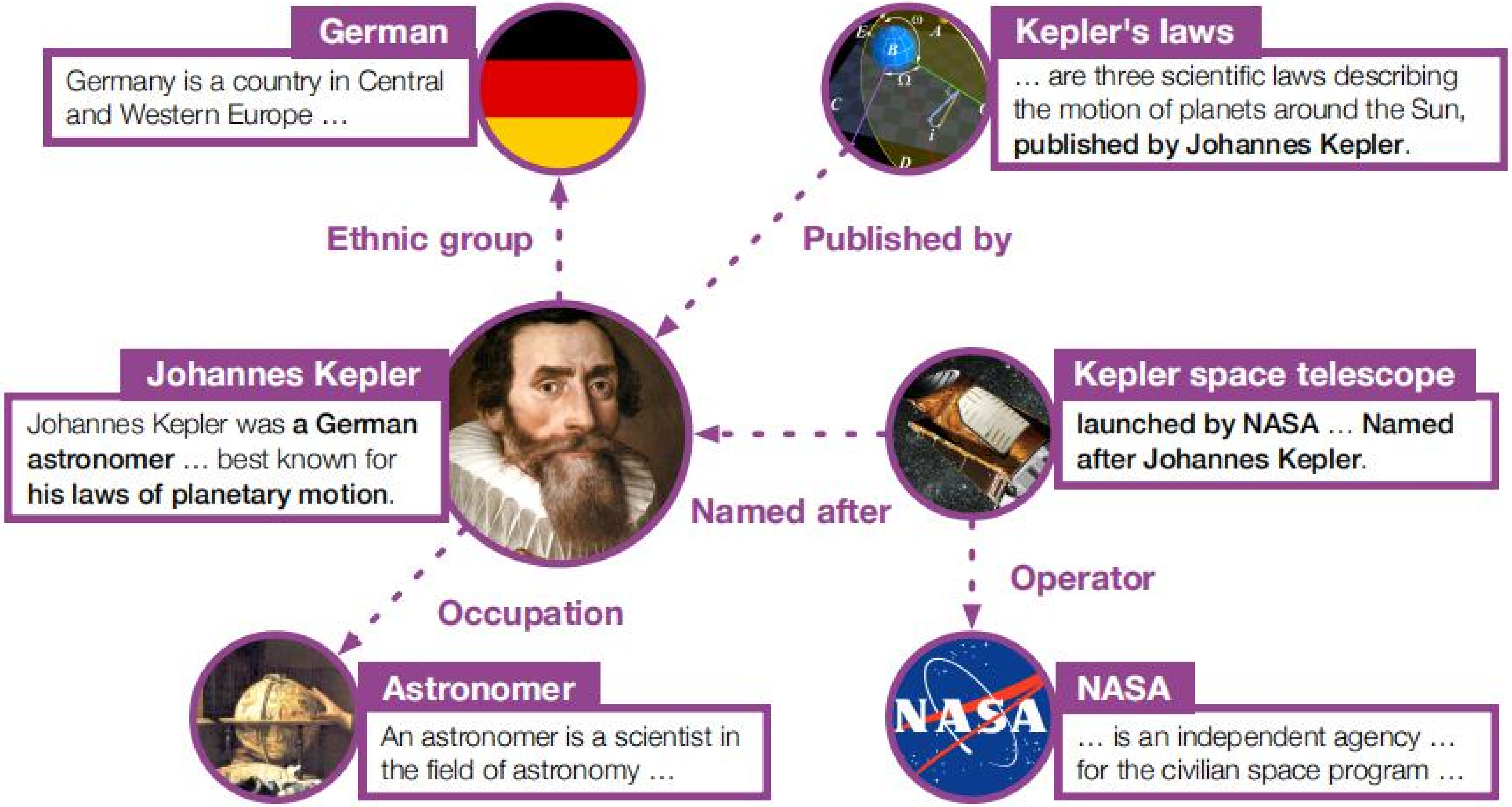

German
Germany is a country in Central and Western Europe …
Kepler's laws
… are three scientific laws describing the motion of planets around the Sun, **published by Johannes Kepler**.
Ethnic group
Published by
Johannes Kepler
Johannes Kepler was **a German astronomer** … best known for **his laws of planetary motion**.
Kepler space telescope
launched by NASA … Named after Johannes Kepler.
Named after
Operator
Occupation
Astronomer
An astronomer is a scientist in the field of astronomy …
NASA
… is an independent agency … for the civilian space program …

Figure 1: An example of a KG with entity descriptions. The figure suggests that descriptions contain abundant information about entities and can help to represent the relational facts between them.

downstream tasks. Usually, there is a special token <s> added to the beginning of the text, and the output at <s> is regarded sentence representation.

2.1 Encoder

For the text encoder, we use Transformer architecture (Vaswani et al., 2017) in the same way as Devlin et al. (2019) and Liu et al. (2019c). The en-

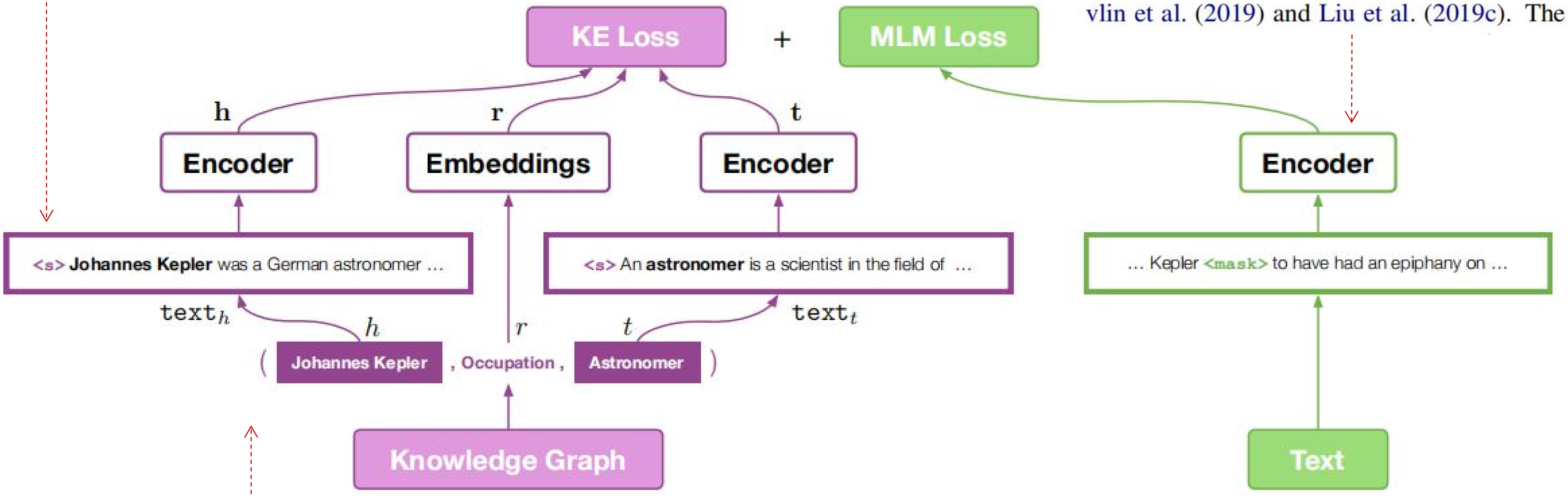
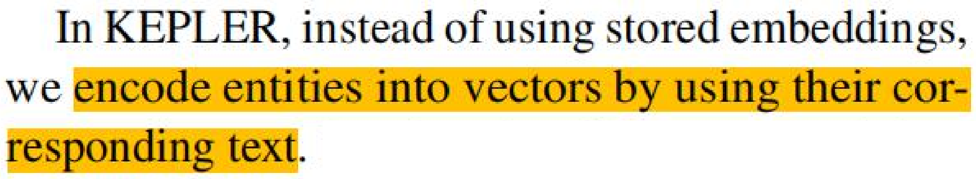

KE Loss
+
MLM Loss
h
r
t
Encoder
Embeddings
Encoder
Encoder
<s> Johannes Kepler was a German astronomer ...
<s> An astronomer is a scientist in the field of ...
... Kepler <mask> to have had an epiphany on ...
text$_h$
h
r
t
text$_t$
(Johannes Kepler , Occupation , Astronomer)
Knowledge Graph
Text

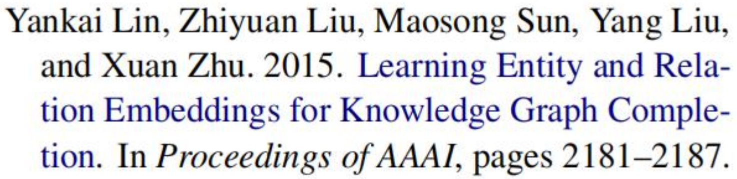
Figure 2: The KEPLER framework. We encode entity descriptions as entity embeddings and jointly train the knowledge embedding (KE) and masked language modeling (MLM) objectives on the same PLM.

The encoder requires a tokenizer to convert plain texts into sequences of tokens. Here we use the same tokenization as RoBERTa: the Byte-Pair Encoding (BPE) (Sennrich et al., 2016).

Unlike previous knowledge-enhanced PLM works (Zhang et al., 2019; Peters et al., 2019), we do not modify the Transformer encoder structure to add external entity linkers or knowledge-integration layers. It means that our model has no



In KEPLER, instead of using stored embeddings, we encode entities into vectors by using their corresponding text.



Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of AAAI*, pages 2181–2187.

- Entity Descriptions as Embeddings

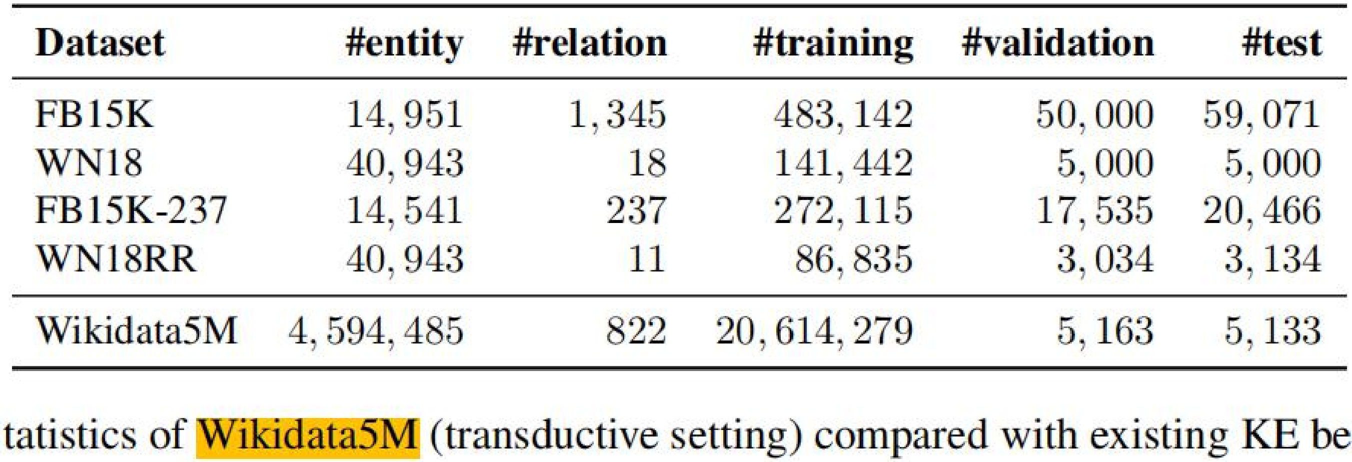
$$\mathbf{h} = \mathtt{E}_{\texttt{<s>}}(\mathtt{text}_h),$$
$$\mathbf{t} = \mathtt{E}_{\texttt{<s>}}(\mathtt{text}_t),$$
$$\mathbf{r} = \mathbf{T}_r,$$

- Entity and Relation Descriptions as Embeddings

$$\hat{\mathbf{r}} = \mathtt{E}_{\texttt{<s>}}(\mathtt{text}_r),$$

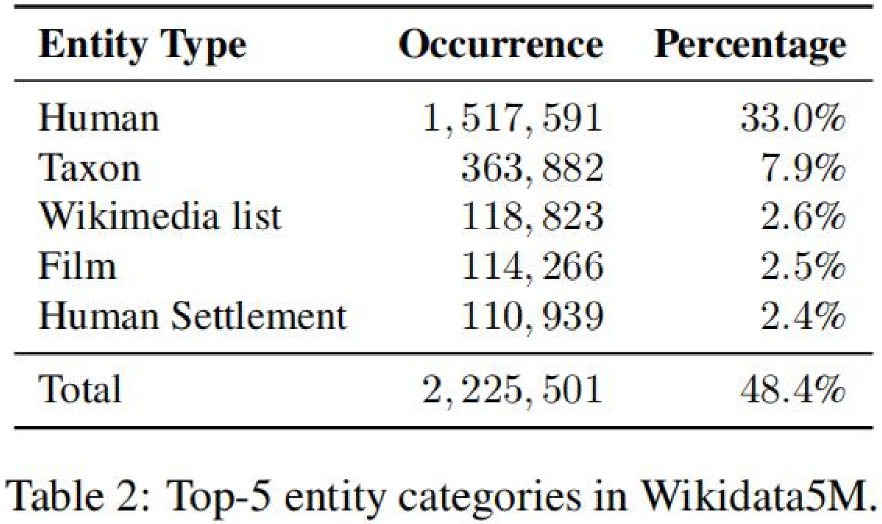
- Entity Embeddings Conditioned on Relations

$$\mathbf{h}_r = \mathtt{E}_{\texttt{<s>}}(\mathtt{text}_{h,r}),$$



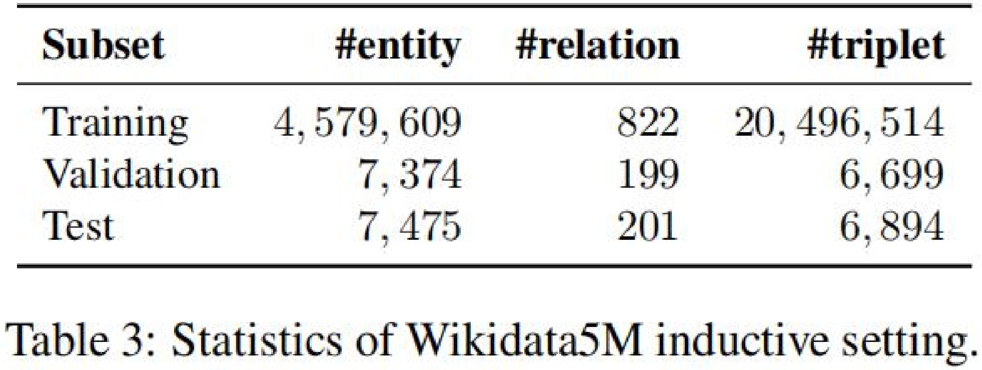
Dataset	#entity	#relation	#training	#validation	#test
FB15K	14,951	1,345	483,142	50,000	59,071
WN18	40,943	18	141,442	5,000	5,000
FB15K-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134
Wikidata5M	4,594,485	822	20,614,279	5,163	5,133

Table 1: Statistics of Wikidata5M (transductive setting) compared with existing KE benchmarks.



Entity Type	Occurrence	Percentage
Human	1, 517, 591	33.0%
Taxon	363, 882	7.9%
Wikimedia list	118, 823	2.6%
Film	114, 266	2.5%
Human Settlement	110, 939	2.4%
Total	2, 225, 501	48.4%

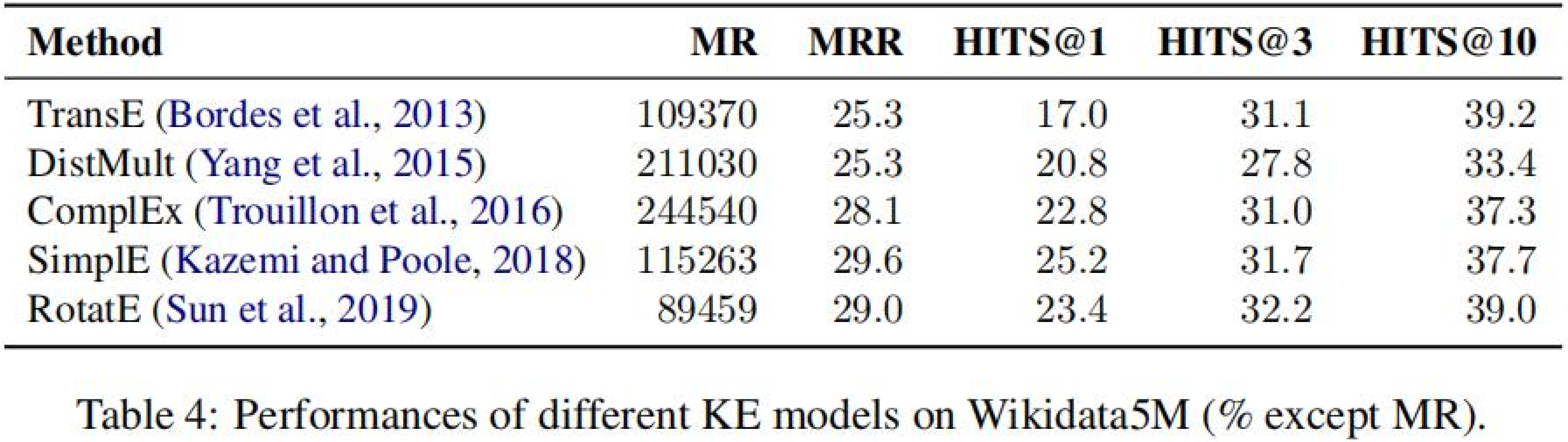
Table 2: Top-5 entity categories in Wikidata5M.



Subset	#entity	#relation	#triplet
Training	4, 579, 609	822	20, 496, 514
Validation	7, 374	199	6, 699
Test	7, 475	201	6, 894

Table 3: Statistics of Wikidata5M inductive setting.

https://www.wikidata.org
https://en.wikipedia.org



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).

Model	FewRel 1.0				FewRel 2.0			
	5-1	**5-5**	**10-1**	**10-5**	**5-1**	**5-5**	**10-1**	**10-5**
MTB (BERT$_{\texttt{LARGE}}$)†	93.86	97.06	89.20	94.27	–	–	–	–
Proto (BERT)	80.68	89.60	71.48	82.89	40.12	51.50	26.45	36.93
Proto (MTB)	81.39	91.05	71.55	83.47	52.13	76.67	48.28	69.75
Proto (ERNIE$_{\texttt{BERT}}$)†	**89.43**	94.66	**84.23**	90.83	49.40	65.55	34.99	49.68
Proto (KnowBert$_{\texttt{BERT}}$)†	86.64	93.22	79.52	88.35	64.40	79.87	51.66	69.71
Proto (RoBERTa)	85.78	95.78	77.65	92.26	64.65	82.76	50.80	71.84
Proto (Our RoBERTa)	84.42	95.30	76.43	91.74	61.98	83.11	48.56	72.19
Proto (ERNIE$_{\texttt{RoBERTa}}$)†	87.76	95.62	80.14	91.47	54.43	80.48	37.97	66.26
Proto (KnowBert$_{\texttt{RoBERTa}}$)†	82.39	93.62	76.21	88.57	55.68	71.82	41.90	58.55
Proto (KEPLER-Wiki)	88.30	**95.94**	81.10	**92.67**	**66.41**	**84.02**	**51.85**	**73.60**
PAIR (BERT)	88.32	93.22	80.63	87.02	**67.41**	78.57	**54.89**	66.85
PAIR (MTB)	83.01	87.64	73.42	78.47	46.18	70.50	36.92	55.17
PAIR (ERNIE$_{\texttt{BERT}}$)†	**92.53**	94.27	**87.08**	89.13	56.18	68.97	43.40	54.35
PAIR (KnowBert$_{\texttt{BERT}}$)†	88.48	92.75	82.57	86.18	66.05	77.88	50.86	67.19
PAIR (RoBERTa)	89.32	93.70	82.49	88.43	66.78	81.84	53.99	70.85
PAIR (Our RoBERTa)	89.26	93.71	83.32	89.02	63.22	77.66	49.28	65.97
PAIR (ERNIE$_{\texttt{RoBERTa}}$)†	87.46	94.11	81.68	87.83	59.29	72.91	48.51	60.26
PAIR (KnowBert$_{\texttt{RoBERTa}}$)†	85.05	91.34	76.04	85.25	50.68	66.04	37.10	51.13
PAIR (KEPLER-Wiki)	90.31	**94.28**	85.48	**90.51**	67.23	**82.09**	54.32	**71.01**

Table 6: Accuracies (%) on the FewRel dataset. N-K indicates the N-way K-shot setting. MTB uses the `LARGE` size and all the other models use the `BASE` size. † indicates oracle models which may have seen facts in the FewRel 1.0 test set during pre-training.

ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang
Baidu Inc., Beijing, China
{sunyu02, wangshuohuan, tianhao, wu_hua, wanghaifeng}@baidu.com

The contributions of this paper:

- They propose a continual pre-training framework ERNIE 2.0, which efficiently supports customized training tasks and continual multi-task learning in an incremental way.
- They construct three kinds of unsupervised language processing tasks to verify the effectiveness of the proposed framework.

- https://github.com/PaddlePaddle/ERNIE.

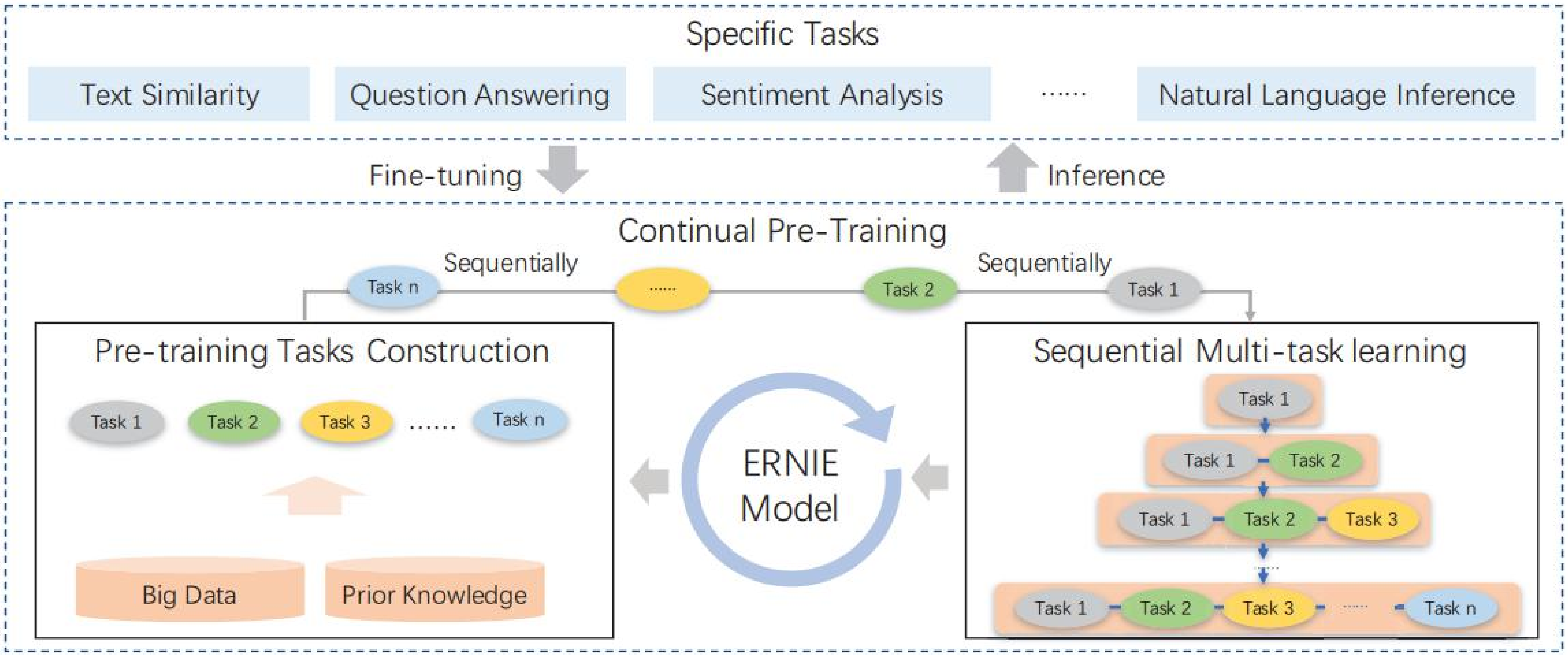


Figure 1: The framework of ERNIE 2.0, where the pre-training tasks can be incrementally constructed, the models are pre-trained through continual multi-task learning, and the pre-trained model is fine-tuned to adapt to various language understanding tasks.

Continual Pre-training **Pre-training Tasks Construction** **Continual Multi-task Learning**

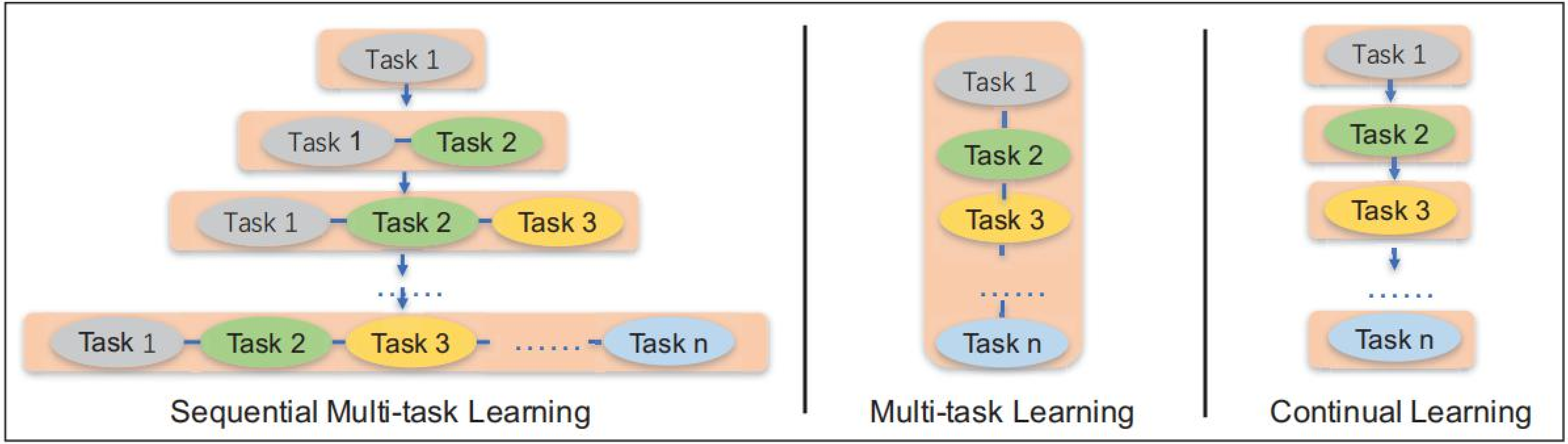
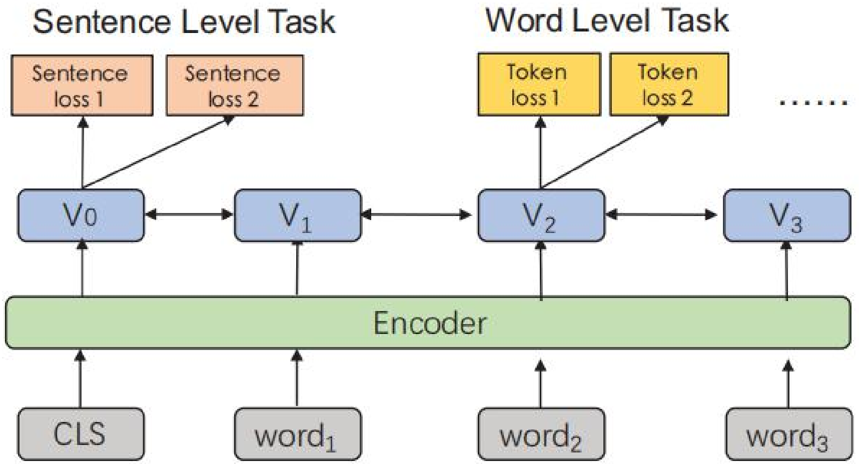


Figure 2: The different methods of continual pre-training.



- **Continual Learning** 训练的每个阶段仅通过一项任务来训练模型。

- **Multi-task Learning** 所有任务一起进行多任务学习。

- **Sequential Multi-task Learning** ERNIE 2.0中新提出的方法，每当有新任务出现时，使用先前学习的参数来初始化模型，并同时训练新引入的任务和原始任务。优点：解决了前两种方法的问题，可随时引入新任务，并保留先前学到的知识。

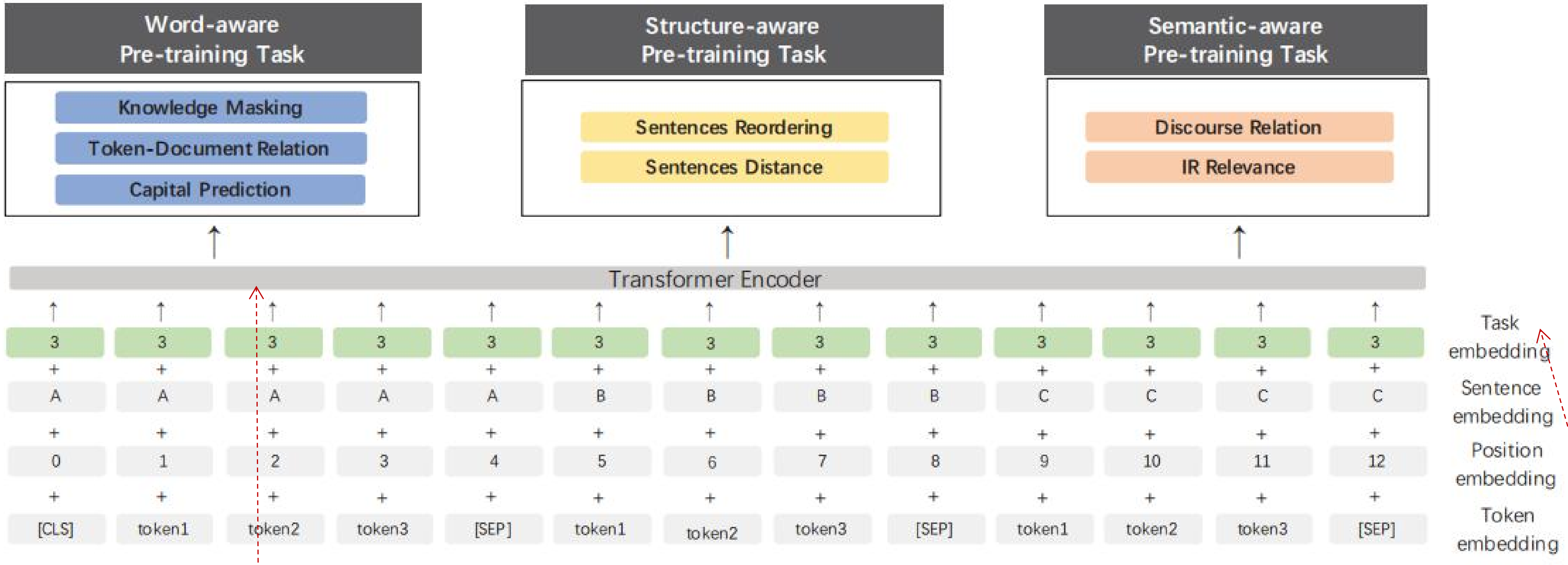
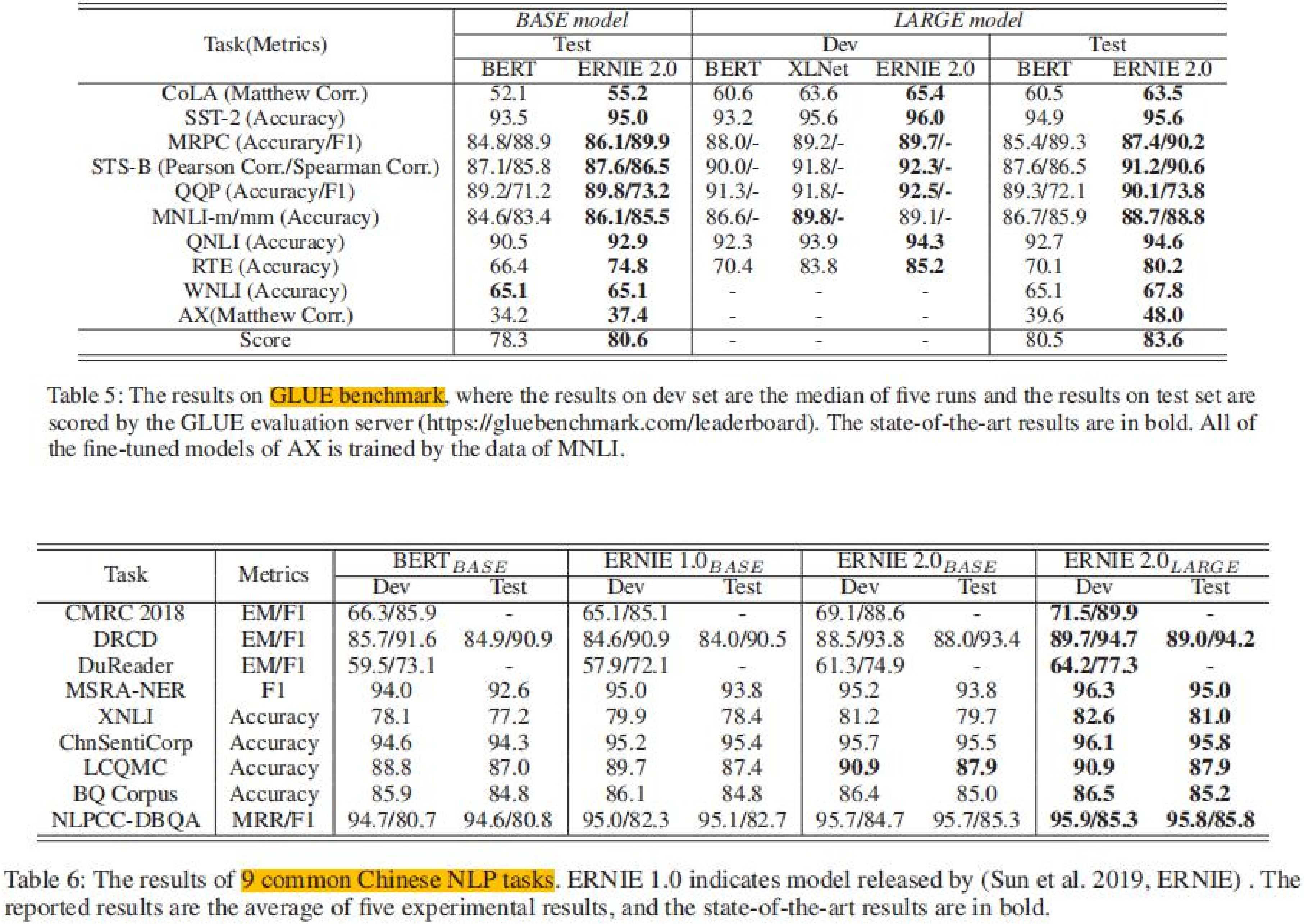


Figure 3: The structure of the ERNIE 2.0 model. The input embedding contains the token embedding, the sentence embedding, the position embedding and the task embedding. Seven pre-training tasks belonging to different kinds are constructed in the ERNIE 2.0 model.

Transformer Encoder The model uses a multi-layer Transformer(Vaswani et al. 2017) as the basic encoder like other pre-training models such as GPT(Radford et al. 2018), BERT(Devlin et al. 2018) and XLM(Lample and Conneau

Task Embedding The model feeds task embedding to represent the characteristic of different tasks. We represent different tasks with an id ranging from 0 to N. Each task id is



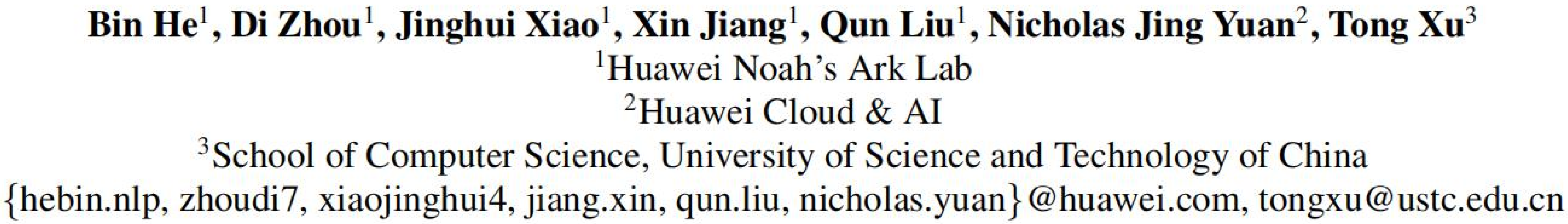
Task(Metrics)	*BASE model*		*LARGE model*				
	Test		Dev			Test	
	BERT	ERNIE 2.0	BERT	XLNet	ERNIE 2.0	BERT	ERNIE 2.0
CoLA (Matthew Corr.)	52.1	**55.2**	60.6	63.6	**65.4**	60.5	**63.5**
SST-2 (Accuracy)	93.5	**95.0**	93.2	95.6	**96.0**	94.9	**95.6**
MRPC (Accurary/F1)	84.8/88.9	**86.1/89.9**	88.0/-	89.2/-	**89.7/-**	85.4/89.3	**87.4/90.2**
STS-B (Pearson Corr./Spearman Corr.)	87.1/85.8	**87.6/86.5**	90.0/-	91.8/-	**92.3/-**	87.6/86.5	**91.2/90.6**
QQP (Accuracy/F1)	89.2/71.2	**89.8/73.2**	91.3/-	91.8/-	**92.5/-**	89.3/72.1	**90.1/73.8**
MNLI-m/mm (Accuracy)	84.6/83.4	**86.1/85.5**	86.6/-	**89.8/-**	89.1/-	86.7/85.9	**88.7/88.8**
QNLI (Accuracy)	90.5	**92.9**	92.3	93.9	**94.3**	92.7	**94.6**
RTE (Accuracy)	66.4	**74.8**	70.4	83.8	**85.2**	70.1	**80.2**
WNLI (Accuracy)	**65.1**	**65.1**	-	-	-	65.1	**67.8**
AX(Matthew Corr.)	34.2	**37.4**	-	-	-	39.6	**48.0**
Score	78.3	**80.6**	-	-	-	80.5	**83.6**

Table 5: The results on GLUE benchmark, where the results on dev set are the median of five runs and the results on test set are scored by the GLUE evaluation server (https://gluebenchmark.com/leaderboard). The state-of-the-art results are in bold. All of the fine-tuned models of AX is trained by the data of MNLI.

Task	Metrics	BERT$_{BASE}$		ERNIE 1.0$_{BASE}$		ERNIE 2.0$_{BASE}$		ERNIE 2.0$_{LARGE}$	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
CMRC 2018	EM/F1	66.3/85.9	-	65.1/85.1	-	69.1/88.6	-	**71.5/89.9**	-
DRCD	EM/F1	85.7/91.6	84.9/90.9	84.6/90.9	84.0/90.5	88.5/93.8	88.0/93.4	**89.7/94.7**	**89.0/94.2**
DuReader	EM/F1	59.5/73.1	-	57.9/72.1	-	61.3/74.9	-	**64.2/77.3**	-
MSRA-NER	F1	94.0	92.6	95.0	93.8	95.2	93.8	**96.3**	**95.0**
XNLI	Accuracy	78.1	77.2	79.9	78.4	81.2	79.7	**82.6**	**81.0**
ChnSentiCorp	Accuracy	94.6	94.3	95.2	95.4	95.7	95.5	**96.1**	**95.8**
LCQMC	Accuracy	88.8	87.0	89.7	87.4	**90.9**	**87.9**	**90.9**	**87.9**
BQ Corpus	Accuracy	85.9	84.8	86.1	84.8	86.4	85.0	**86.5**	**85.2**
NLPCC-DBQA	MRR/F1	94.7/80.7	94.6/80.8	95.0/82.3	95.1/82.7	95.7/84.7	95.7/85.3	**95.9/85.3**	**95.8/85.8**

Table 6: The results of 9 common Chinese NLP tasks. ERNIE 1.0 indicates model released by (Sun et al. 2019, ERNIE) . The reported results are the average of five experimental results, and the state-of-the-art results are in bold.

BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models

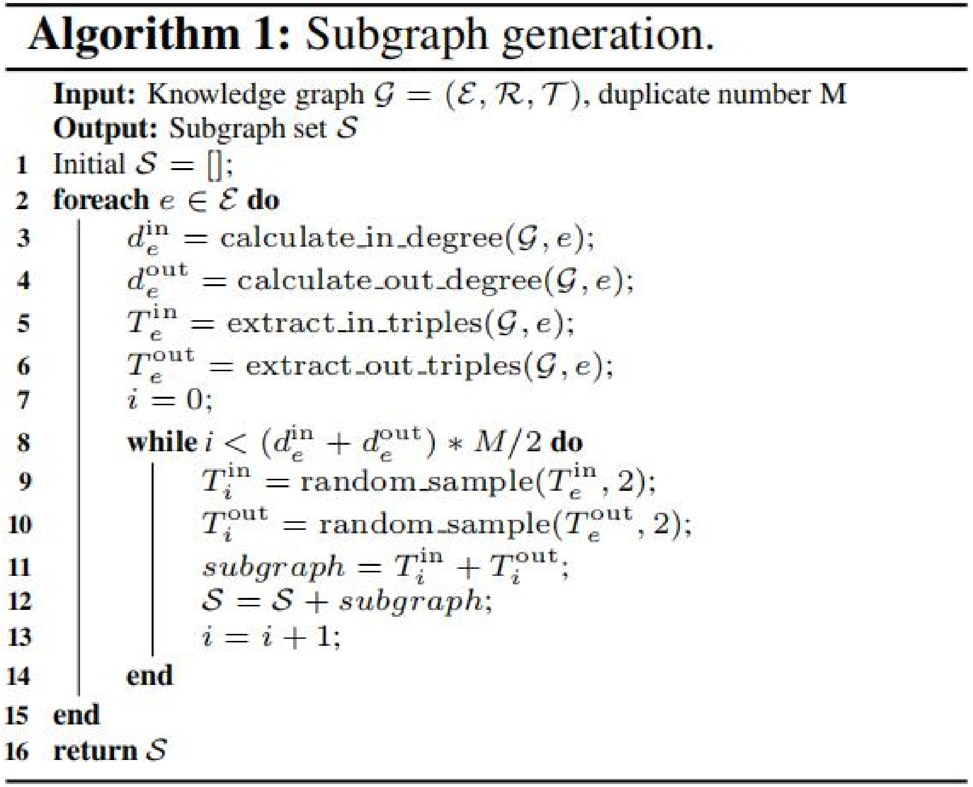


Bin He[1], Di Zhou[1], Jinghui Xiao[1], Xin Jiang[1], Qun Liu[1], Nicholas Jing Yuan[2], Tong Xu[3]
[1]Huawei Noah's Ark Lab
[2]Huawei Cloud & AI
[3]School of Computer Science, University of Science and Technology of China
{hebin.nlp, zhoudi7, xiaojinghui4, jiang.xin, qun.liu, nicholas.yuan}@huawei.com, tongxu@ustc.edu.cn

The contributions of this paper:

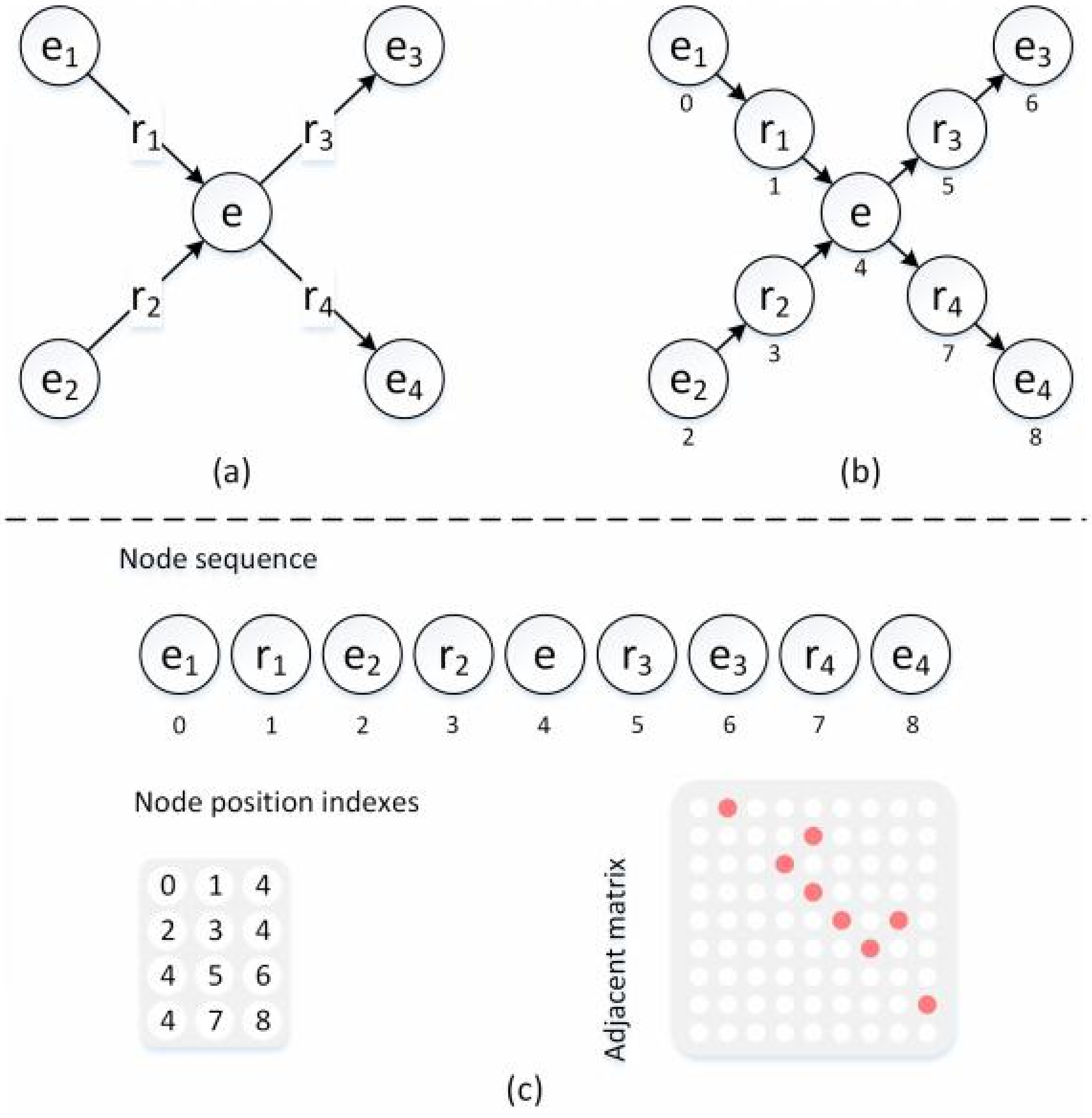
- We propose a novel knowledge-enhanced pretrained language model BERT-MK for medical NLP tasks, which integrates graph contextualized knowledge learned from the medical KG.

- Experimental results show that BERT-MK achieves better performance than previous state-of-the-art biomedical pre-trained language models on entity typing and relation classification tasks.



Algorithm 1: Subgraph generation.

```
Input: Knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, duplicate number M
Output: Subgraph set $\mathcal{S}$
Initial $\mathcal{S} = []$;
foreach $e \in \mathcal{E}$ do
    $d_e^{in}$ = calculate_in_degree($\mathcal{G}, e$);
    $d_e^{out}$ = calculate_out_degree($\mathcal{G}, e$);
    $T_e^{in}$ = extract_in_triples($\mathcal{G}, e$);
    $T_e^{out}$ = extract_out_triples($\mathcal{G}, e$);
    $i = 0$;
    while $i < (d_e^{in} + d_e^{out}) * M/2$ do
        $T_i^{in}$ = random_sample($T_e^{in}$, 2);
        $T_i^{out}$ = random_sample($T_e^{out}$, 2);
        $subgraph = T_i^{in} + T_i^{out}$;
        $\mathcal{S} = \mathcal{S} + subgraph$;
        $i = i + 1$;
    end
end
return $\mathcal{S}$
```



(a)

(b)

Node sequence

Node position indexes

0	1	4
2	3	4
4	5	6
4	7	8

(c)

Downstream Tasks

Entity Typing

Relation Classification

Baselines

BERT-Base (Devlin et al., 2019)

BioBERT (Lee et al., 2019)

SCIBERT (Beltagy et al., 2019)

Dataset

https://www.ncbi.nlm.nih.gov/pubmed/

https://www.ncbi.nlm.nih.gov/pmc/

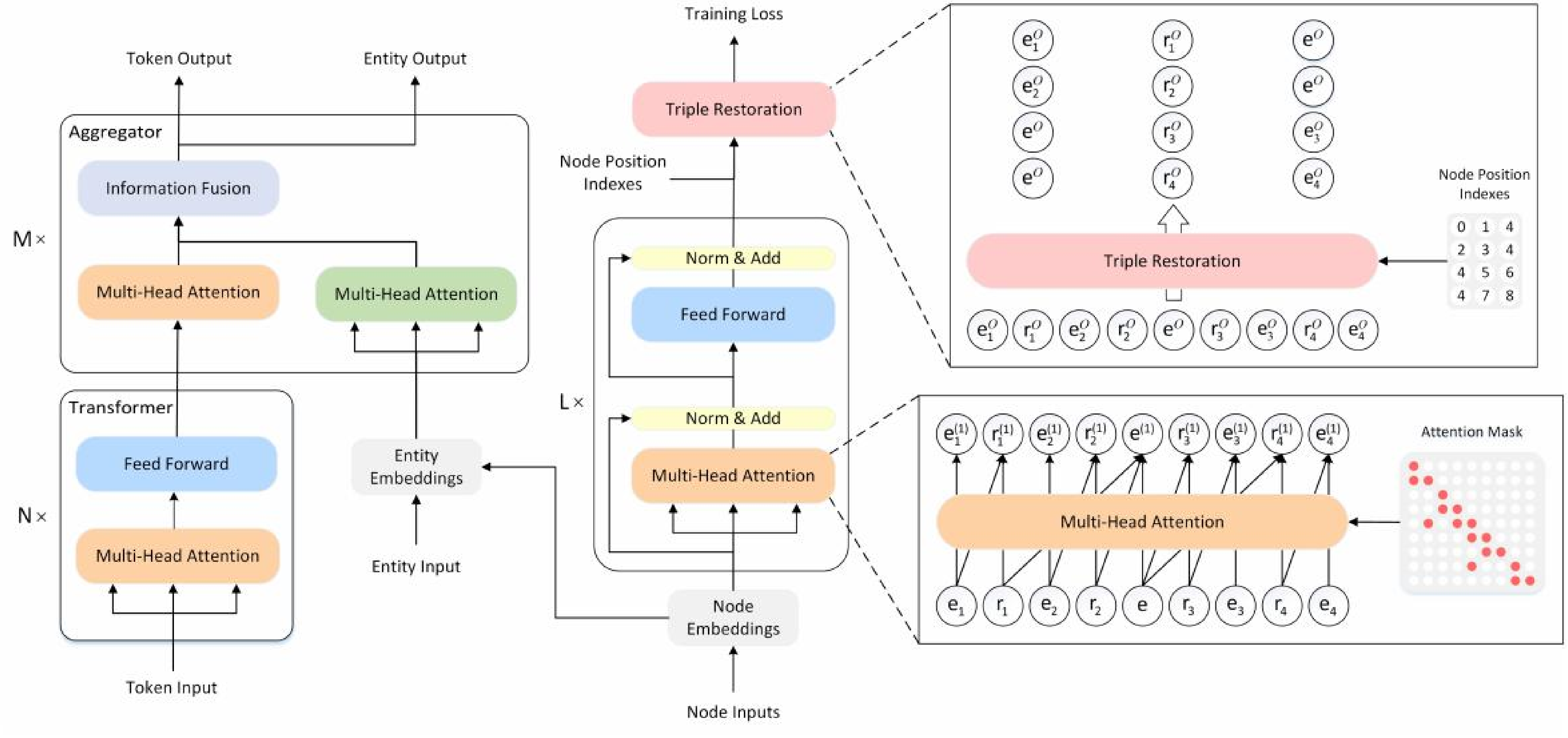


Figure 3: The model architecture of BERT-MK. The left part is the pre-trained language model, in which entity information learned from the knowledge graph is incorporated. The right part is GCKE module. The subgraph in Figure 2 is utilized to describe the learning process. e_1, $e_1^{(1)}$, e_1^O is the embedding of the input node, the updated node and the output node, respectively.

Table 2: Statistics of the datasets. Most of these datasets do not follow a standard train-valid-test set partition, and we adopt some traditional data partition ways to do model training and evaluation.

Task	Dataset	# Train	# Valid	# Test
Entity Typing	2010 i2b2/VA (Uzuner et al., 2011)	16,519	-	31,161
	JNLPBA (Kim et al., 2004)	51,301	-	8,653
	BC5CDR (Li et al., 2016)	9,385	9,593	9,809
Relation Classification	2010 i2b2/VA (Uzuner et al., 2011)	10,233	-	19,115
	GAD (Bravo et al., 2015)	5,339	-	-
	EU-ADR (Van Mulligen et al., 2012)	355	-	-

Task	Dataset	Metrics	E-SVM	CNN-M	BERT-Base	BioBERT	SCIBERT	BERT-MK
Entity Typing	2010 i2b2/VA	Acc	-	-	96.76	97.43	**97.74**	97.70
	JNLPBA	Acc	-	-	94.12	94.37	**94.60**	94.55
	BC5CDR	Acc	-	-	98.78	99.27	99.38	**99.54**
Relation Classification	2010 i2b2/VA	P	-	73.1	72.6	76.1	74.8	**77.6**
		R	-	66.7	65.7	71.3	71.6	**72.0**
		F	-	69.7	69.2	73.6	73.1	**74.7**
	GAD	P	79.21	-	74.28	76.43	77.47	**81.67**
		R	89.25	-	85.11	87.65	85.94	**92.79**
		F	83.93	-	79.33	81.66	81.45	**86.87**
	EU-ADR	P	-	-	75.45	81.05	78.42	**84.43**
		R	-	-	**96.55**	93.90	90.09	91.17
		F	-	-	84.71	87.00	85.51	**87.49**

Thanks