BERT + Knowledge

Немного о KR

Translating Embeddings for Modeling Multi-relational Data

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Dura´n, Jason Weston, Oksana Yakhnenko, 2013

$$\mathcal{L} = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$$

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$

  $\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$

  $\mathbf{e} \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$

loop

  $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$

  $S_{batch} \leftarrow$sample(S, b) // sample a minibatch of size b

  $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets

  for $(h, \ell, t) \in S_{batch}$ **do**

    $(h', \ell, t') \leftarrow$sample$(S'_{(h,\ell,t)})$ // sample a corrupted triplet

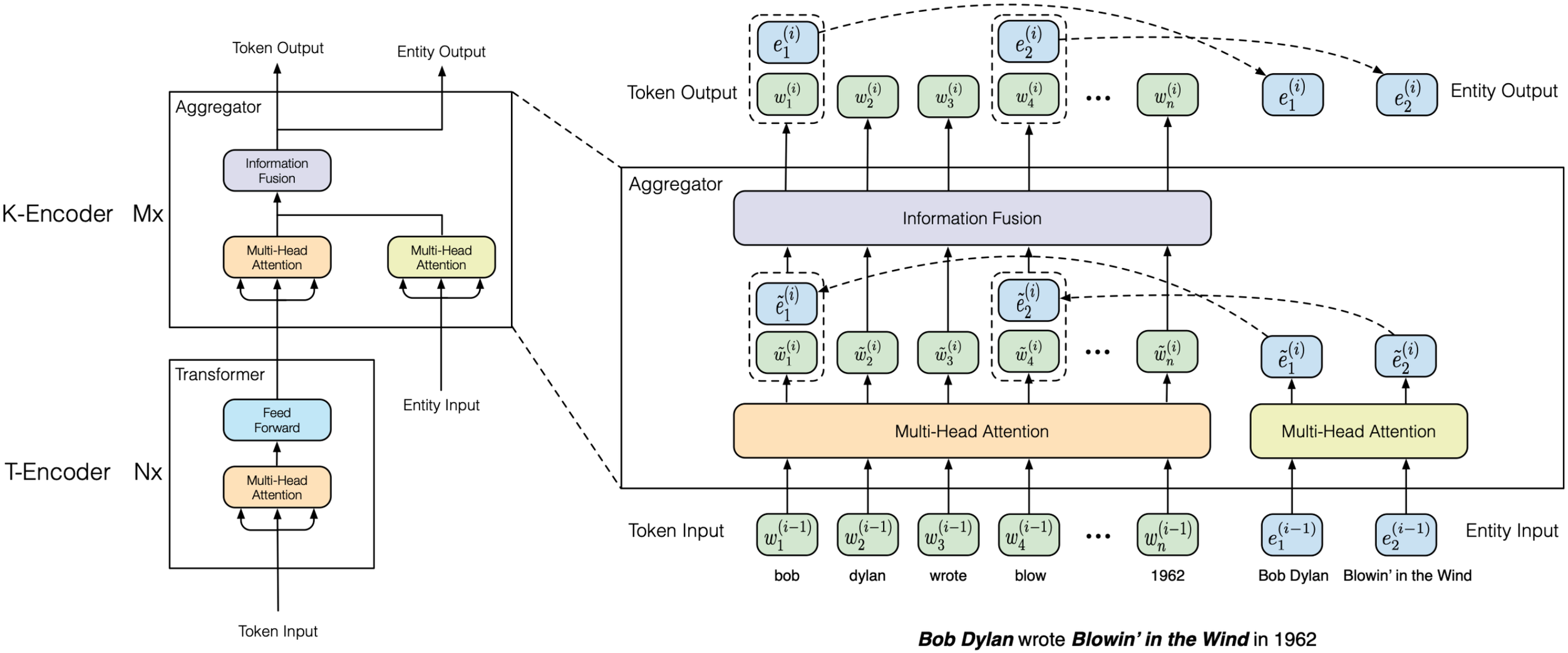
    $T_{batch} \leftarrow T_{batch} \cup \left\{\left((h, \ell, t), (h', \ell, t')\right)\right\}$

  end for

  Update embeddings w.r.t. $\sum_{\left((h,\ell,t),(h',\ell,t')\right)\in T_{batch}} \nabla\left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h'}+\boldsymbol{\ell},\boldsymbol{t'})\right]_+$

end loop

ERNIE: Enhanced Language Representation with Informative Entities



(a) Model Achitecture

(b) Aggregator

$$\{\tilde{\boldsymbol{w}}_1^{(i)}, \ldots, \tilde{\boldsymbol{w}}_n^{(i)}\} = \texttt{MH-ATT}(\{\boldsymbol{w}_1^{(i-1)}, \ldots, \boldsymbol{w}_n^{(i-1)}\}),$$
$$\{\tilde{\boldsymbol{e}}_1^{(i)}, \ldots, \tilde{\boldsymbol{e}}_m^{(i)}\} = \texttt{MH-ATT}(\{\boldsymbol{e}_1^{(i-1)}, \ldots, \boldsymbol{e}_m^{(i-1)}\}).$$

$$\boldsymbol{h}_j = \sigma(\tilde{\boldsymbol{W}}_t^{(i)} \tilde{\boldsymbol{w}}_j^{(i)} + \tilde{\boldsymbol{W}}_e^{(i)} \tilde{\boldsymbol{e}}_k^{(i)} + \tilde{\boldsymbol{b}}^{(i)}),$$
$$\boldsymbol{w}_j^{(i)} = \sigma(\boldsymbol{W}_t^{(i)} \boldsymbol{h}_j + \boldsymbol{b}_t^{(i)}),$$
$$\boldsymbol{e}_k^{(i)} = \sigma(\boldsymbol{W}_e^{(i)} \boldsymbol{h}_j + \boldsymbol{b}_e^{(i)}).$$

$$\boldsymbol{h}_j = \sigma(\tilde{\boldsymbol{W}}_t^{(i)} \tilde{\boldsymbol{w}}_j^{(i)} + \tilde{\boldsymbol{b}}^{(i)}),$$
$$\boldsymbol{w}_j^{(i)} = \sigma(\boldsymbol{W}_t^{(i)} \boldsymbol{h}_j + \boldsymbol{b}_t^{(i)}).$$

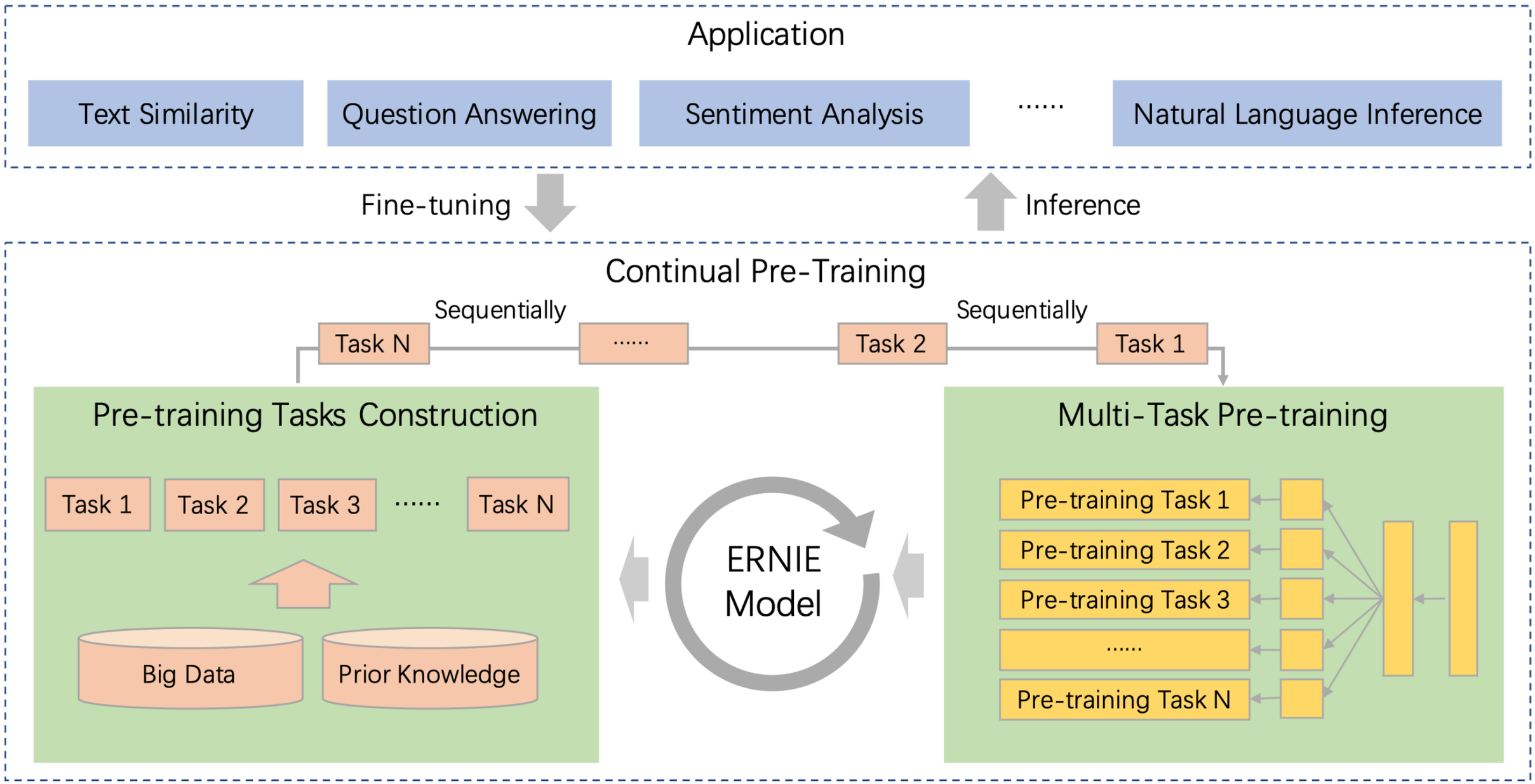
$$p(e_j|w_i) = \frac{\exp(\texttt{linear}(\boldsymbol{w}_i^o) \cdot \boldsymbol{e}_j)}{\sum_{k=1}^{m} \exp(\texttt{linear}(\boldsymbol{w}_i^o) \cdot \boldsymbol{e}_k)},$$

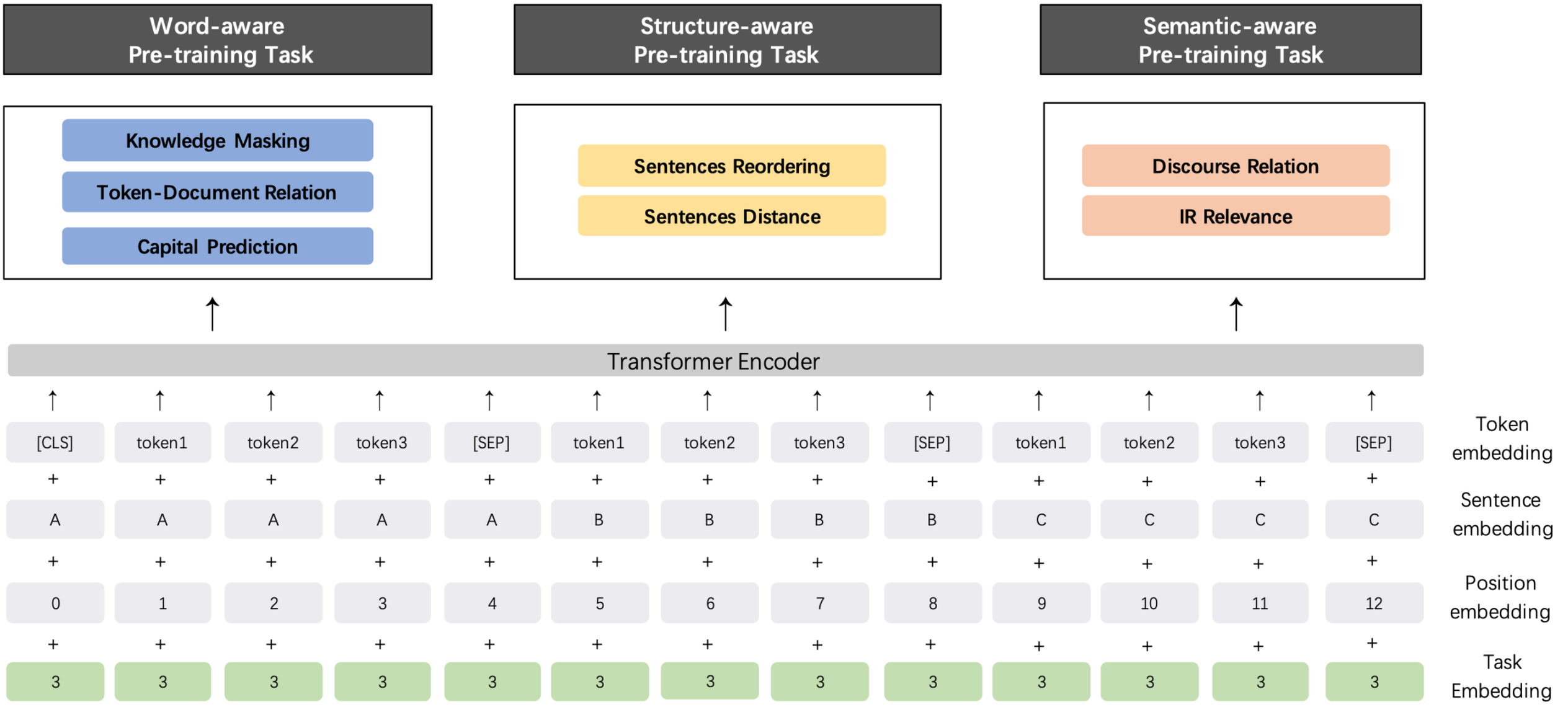
ERNIE: Enhanced Representation through Knowledge Integration

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

No	Text	Predict by ERNIE	Predict by BERT	Answer
1	2006年9月，_____与张柏芝结婚，两人婚后育有两儿子——大儿子Lucas谢振轩，小儿子Quintus谢振南；	谢霆锋	谢振轩	谢霆锋
	In September 2006, _____ married Cecilia Cheung. They had two sons, the older one is Zhenxuan Xie and the younger one is Zhennan Xie.	Tingfeng Xie	Zhenxuan Xie	Tingfeng Xie
2	戊戌变法，又称百日维新，是_____、梁启超等维新派人士通过光绪帝进行的一场资产阶级改良。	康有为	孙世昌	康有为
	The Reform Movement of 1898, also known as the Hundred-Day Reform, was a bourgeois reform carried out by the reformists such as ____ and Qichao Liang through Emperor Guangxu.	Youwei Kang	Shichang Sun	Youwei Kang
3	高血糖则是由于_____分泌缺陷或其生物作用受损，或两者兼有引起。糖尿病时长期存在的高血糖，导致各种组织，特别是眼、肾、心脏、血管、神经的慢性损害、功能障碍。	胰岛素	糖糖内	胰岛素
	Hyperglycemia is caused by defective _____ secretion or impaired biological function, or both. Long-term hyperglycemia in diabetes leads to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves.	Insulin	(Not a word in Chinese)	Insulin
4	澳大利亚是一个高度发达的资本主义国家，首都为_____。作为南半球经济最发达的国家和全球第12大经济体、全球第四大农产品出口国，其也是多种矿产出口量全球第一的国家。	墨尔本	墨悉本	堪培拉
	Australia is a highly developed capitalist country with _____ as its capital. As the most developed country in the Southern Hemisphere, the 12th largest economy in the world and the fourth largest exporter of agricultural products in the world, it is also the world's largest exporter of various minerals.	Melbourne	(Not a city name)	Canberra (the capital of Australia)
5	_____是中国神魔小说的经典之作，达到了古代长篇浪漫主义小说的巅峰，与《三国演义》《水浒传》《红楼梦》并称为中国古典四大名著。	西游记	《小》	西游记
	_____is a classic novel of Chinese gods and demons, which reaching the peak of ancient Romantic novels. It is also known as the four classical works of China with Romance of the Three Kingdoms, Water Margin and Dream of Red Mansions.	The Journey to the West	(Not a word in Chinese)	The Journey to the West
6	相对论是关于时空和引力的理论，主要由_____创立。	爱因斯坦	卡尔斯所	爱因斯坦
	Relativity is a theory about space-time and gravity, which was founded by _____.	Einstein	(Not a word in Chinese)	Einstein

ERNIE 2.0：A Continual Pre-training framework for Language Understanding




Word-aware Pre-training Task
Structure-aware Pre-training Task
Semantic-aware Pre-training Task
Knowledge Masking
Token-Document Relation
Capital Prediction
Sentences Reordering
Sentences Distance
Discourse Relation
IR Relevance
Transformer Encoder
[CLS] token1 token2 token3 [SEP] token1 token2 token3 [SEP] token1 token2 token3 [SEP]
A A A A A B B B B C C C C
0 1 2 3 4 5 6 7 8 9 10 11 12
3 3 3 3 3 3 3 3 3 3 3 3 3
Token embedding
Sentence embedding
Position embedding
Task Embedding

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**

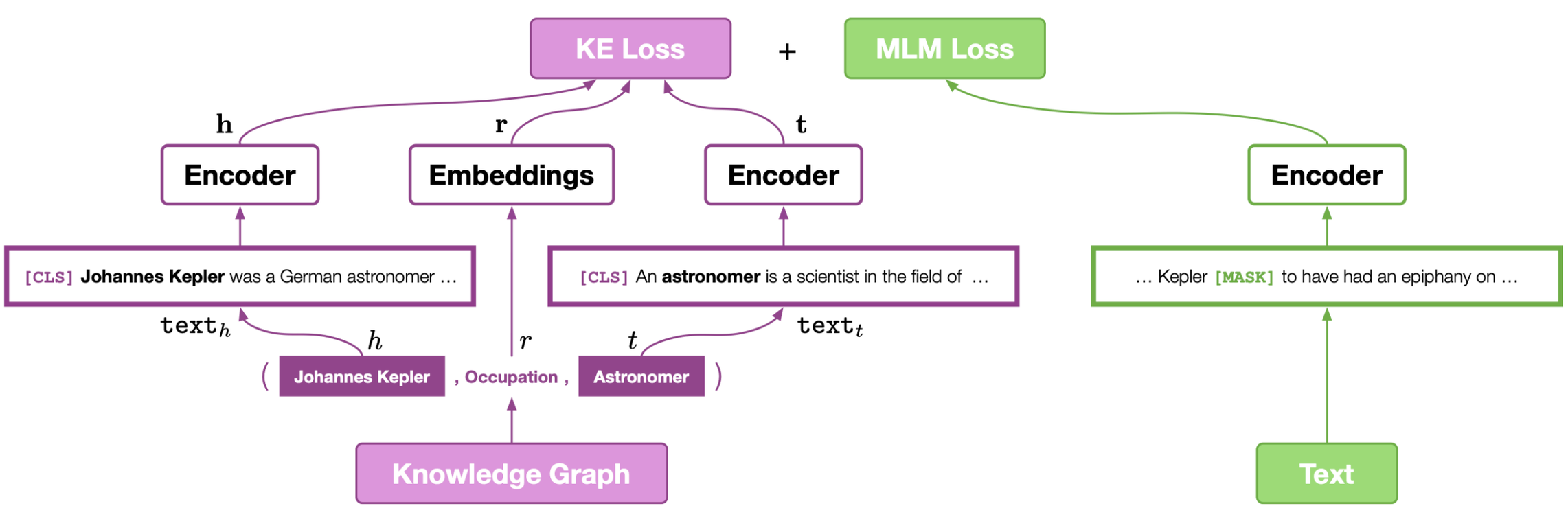
KEPLER

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

$$\mathcal{L}_{\mathrm{KE}} = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} \frac{1}{n} \log \sigma(d_r(\mathbf{h_i'}, \mathbf{t_i'}) - \gamma),$$

KEPLER

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



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 (%).

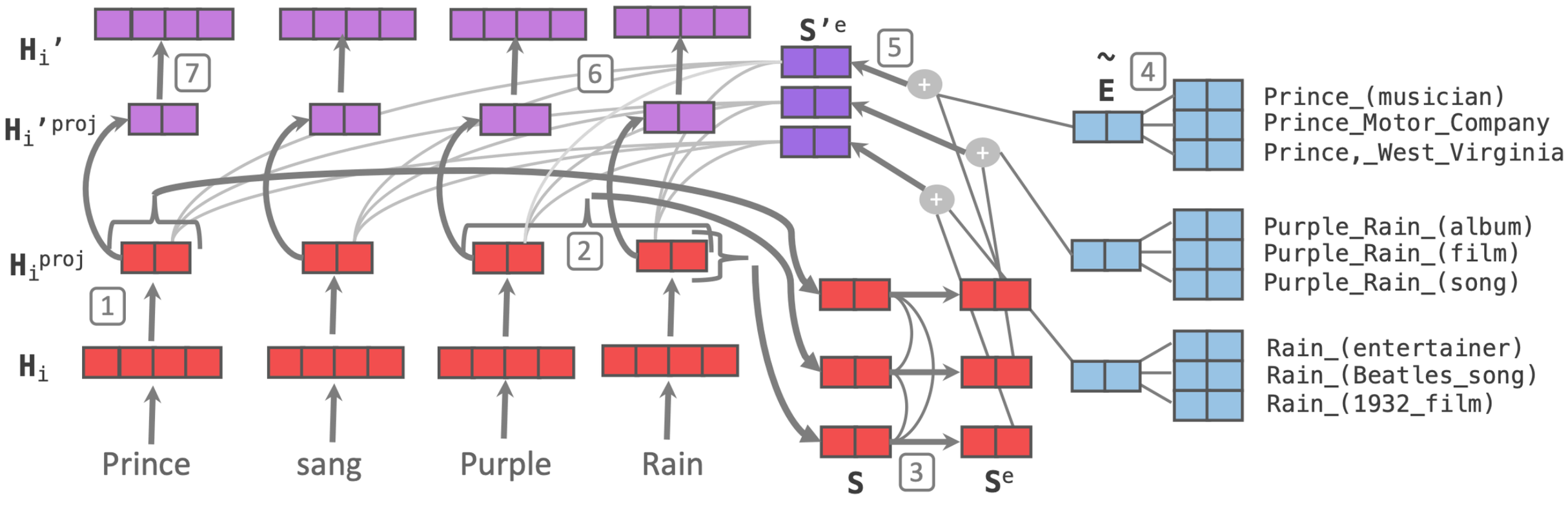
Model	P	R	F-1
BERT	67.2	64.8	66.0
$\text{BERT}_{\text{LARGE}}$	-	-	70.1
RoBERTa	71.1	70.5	70.8
ERNIE	70.0	66.1	68.0
MTB	-	-	71.5
KnowBERT	71.6	71.4	71.5
RoBERTa*	71.3	69.8	70.5
KEPLER-Wiki	72.8	72.2	**72.5**
KEPLER-WordNet	73.0	69.3	71.1
KEPLER-W+W	72.5	72.1	72.3

KnowBert

Matthew E. Peters[1], Mark Neumann[1], Robert L. Logan IV[2], Roy Schwartz[1,3],

Vidur Joshi[1], Sameer Singh[2], and Noah A. Smith[1,3]

9 sep 2019



TuckER: Tensor factorization for knowledge graph completion

Model	Scoring Function	Relation Parameters	Space Complexity
RESCAL (Nickel et al., 2011)	$\mathbf{e}_s^\top \mathbf{W}_r \mathbf{e}_o$	$\mathbf{W}_r \in \mathbb{R}^{{d_e}^2}$	$\mathcal{O}(n_e d_e + n_r d_r^2)$
DistMult (Yang et al., 2015)	$\langle \mathbf{e}_s, \mathbf{w}_r, \mathbf{e}_o \rangle$	$\mathbf{w}_r \in \mathbb{R}^{d_e}$	$\mathcal{O}(n_e d_e + n_r d_e)$
ComplEx (Trouillon et al., 2016)	$\text{Re}(\langle \mathbf{e}_s, \mathbf{w}_r, \overline{\mathbf{e}}_o \rangle)$	$\mathbf{w}_r \in \mathbb{C}^{d_e}$	$\mathcal{O}(n_e d_e + n_r d_e)$
ConvE (Dettmers et al., 2018)	$f(\text{vec}(f([\underline{\mathbf{e}}_s; \underline{\mathbf{w}}_r] * w))\mathbf{W})\mathbf{e}_o$	$\mathbf{w}_r \in \mathbb{R}^{d_r}$	$\mathcal{O}(n_e d_e + n_r d_r)$
SimplE (Kazemi and Poole, 2018)	$\frac{1}{2}(\langle \mathbf{h}_{e_s}, \mathbf{w}_r, \mathbf{t}_{e_o} \rangle + \langle \mathbf{h}_{e_o}, \mathbf{w}_{r^{-1}}, \mathbf{t}_{e_s} \rangle)$	$\mathbf{w}_r \in \mathbb{R}^{d_e}$	$\mathcal{O}(n_e d_e + n_r d_e)$
HypER (Balažević et al., 2019)	$f(\text{vec}(\mathbf{e}_s * \text{vec}^{-1}(\mathbf{w}_r \mathbf{H}))\mathbf{W})\mathbf{e}_o$	$\mathbf{w}_r \in \mathbb{R}^{d_r}$	$\mathcal{O}(n_e d_e + n_r d_r)$
TuckER (ours)	$\mathcal{W} \times_1 \mathbf{e}_s \times_2 \mathbf{w}_r \times_3 \mathbf{e}_o$	$\mathbf{w}_r \in \mathbb{R}^{d_r}$	$\mathcal{O}(n_e d_e + n_r d_r)$

Towards AGI