基于知识增强的语言表示模型

清华大学　张正彦, 韩旭

June 15, 2019

概览

基于特征的预训练语言模型

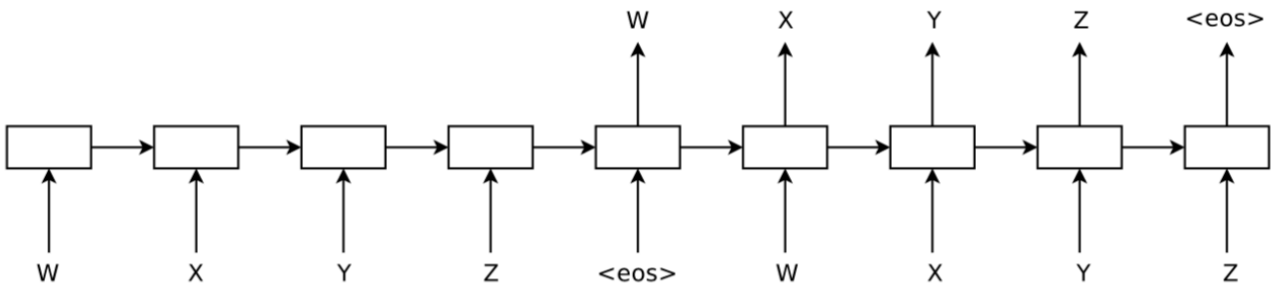
ELMo: Deep contextualized word representations (NAACL 18)

- **早期的 Word2Vec、Glove 得到的特征被广泛用于各类 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**

基于微调的预训练语言模型

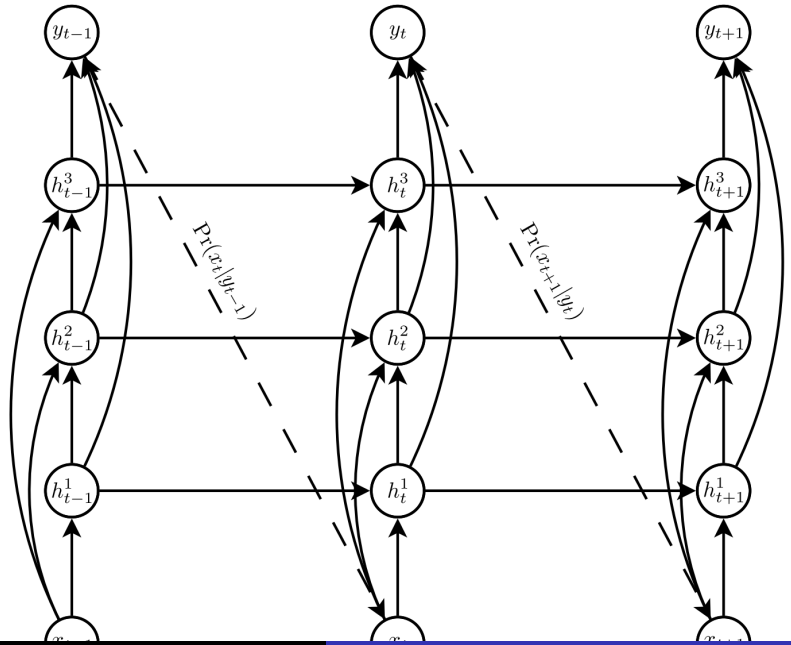
Semi-supervised Sequence Learning (NIPS 15)

- **核心思路：利用大量的无标注数据训练循环神经网络，并在无监督训练得到的循环神经网络上微调，以适应下游的序列学习任务**
- **方案一：采用 seq2seq 模型类似的 Sequence Autoencoder 来对无标注数据进行语言模型的学习**



Semi-supervised Sequence Learning (NIPS 15)

- 核心思路：利用大量的无标注数据训练循环神经网络，并在无监督训练得到的循环神经网络上微调，以适应下游的序列学习任务
- 方案二：采用基于循环神经网络的语言模型来对无标注数据进行学习

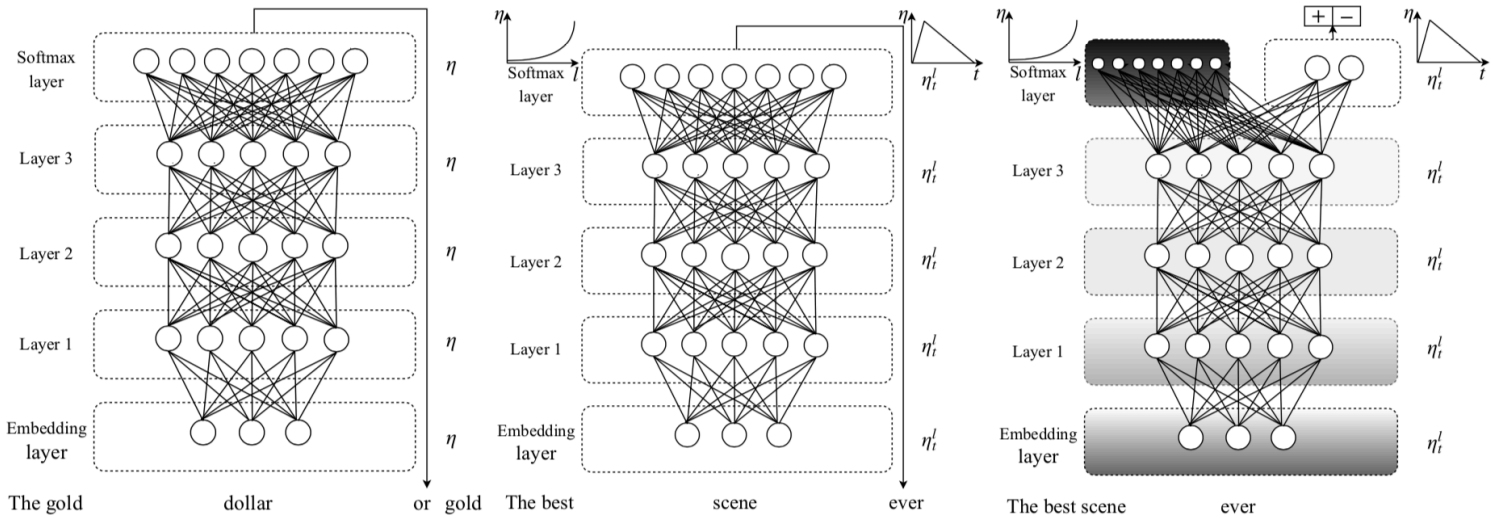


Semi-supervised Sequence Learning (NIPS 15)

- 在情感分析以及文本分类上，将两种预训练方案得到的模型参数作为后续任务模型的初始化参数
- 模型的鲁棒性与效果均得到了显著提升
- 验证了在无监督语料学习参数的基础上，通过“微调”来在下游任务上应用的可能性
- 但该模型对领域数据（in-domain）的需要以及在部分数据上的过拟合仍然需要解决

ULMFiT: Universal Language Model Fine-tuning for Text Classification (ACL 18)

- **核心思路：在大量的通用数据（general-domain）上预训练语言模型，并微调预训练的参数以适应下游目标任务**
- **分为三个阶段：预训练、语言模型微调、目标任务微调**



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

ULMFiT: Universal Language Model Fine-tuning for Text Classification (ACL 18)

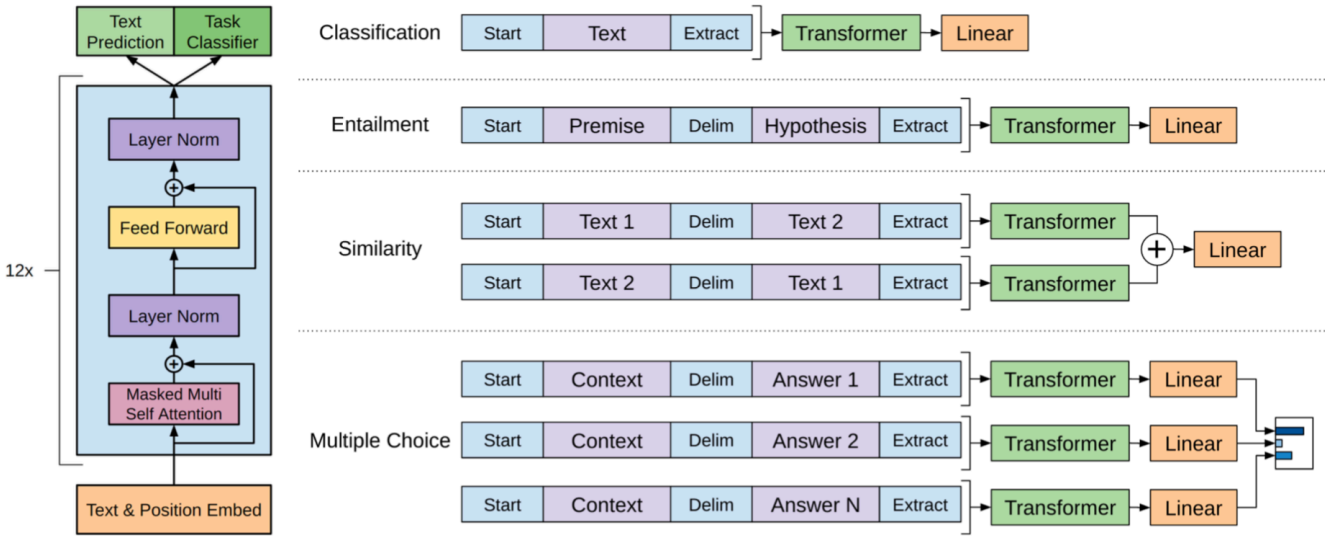
- 在预训练阶段和语言模型微调阶段：
- 采用 AWD-LSTM，能进行语言模型学习的强大模型
- 不同层设置不同的学习率，底层通用性特征设置小学习率，高层具体任务相关特征设置大学习率
- 采用倾斜三角学习率（STLR），先线性增加学习率，然后线性衰减
- 这三点或多或少都被后续模型采用

ULMFiT: Universal Language Model Fine-tuning for Text Classification (ACL 18)

- 在目标任务微调阶段：
- 在语言模型后添加网络来进行文本分类任务
- 分别训练了前向和后向的语言模型，在微调阶段取平均结果来预测
- 采用了逐层解冻（gradual unfreezing）的方法，先微调最后一层，再微调倒数第二层和最后一层，以此类推，同时兼顾收敛效率与平稳

GPT: Improving Language Understanding by Generative Pre-Training

- **采用了 Transformer 代替 LSTM 在无监督数据上预训练语言模型**
- **针对不同任务采取不同的微调模式**

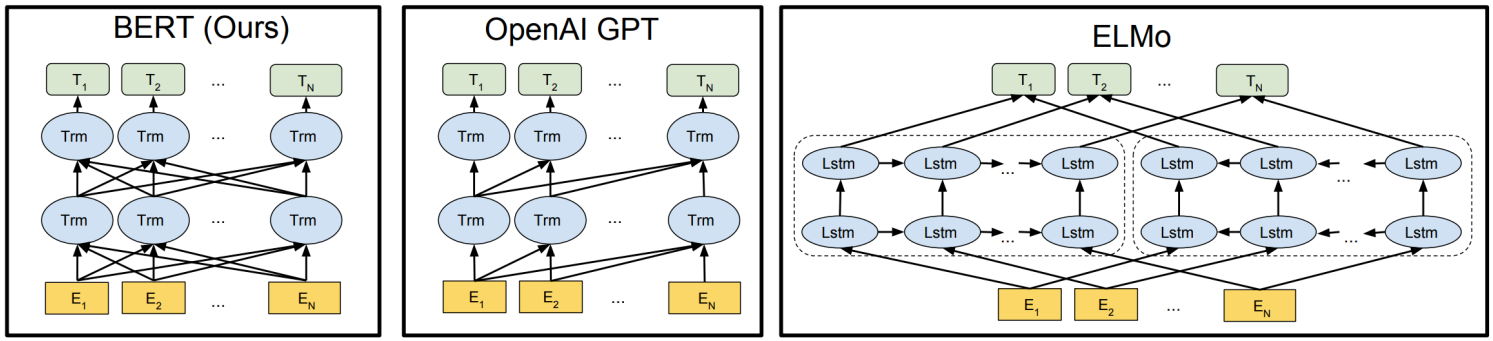


GPT-2: Language Models are Unsupervised Multitask Learners

- 更大数据带来了更好的效果
- 更大模型带来了更好的效果
- 语言模型本身有点做无监督多任务学习的意味，下游的有监督任务某种程度上是语言模型训练下的子任务，这意味着预训练采用更多训练任务也能带来提升
- GPT-2 在文本生成上效果显著（too dangerous to release）

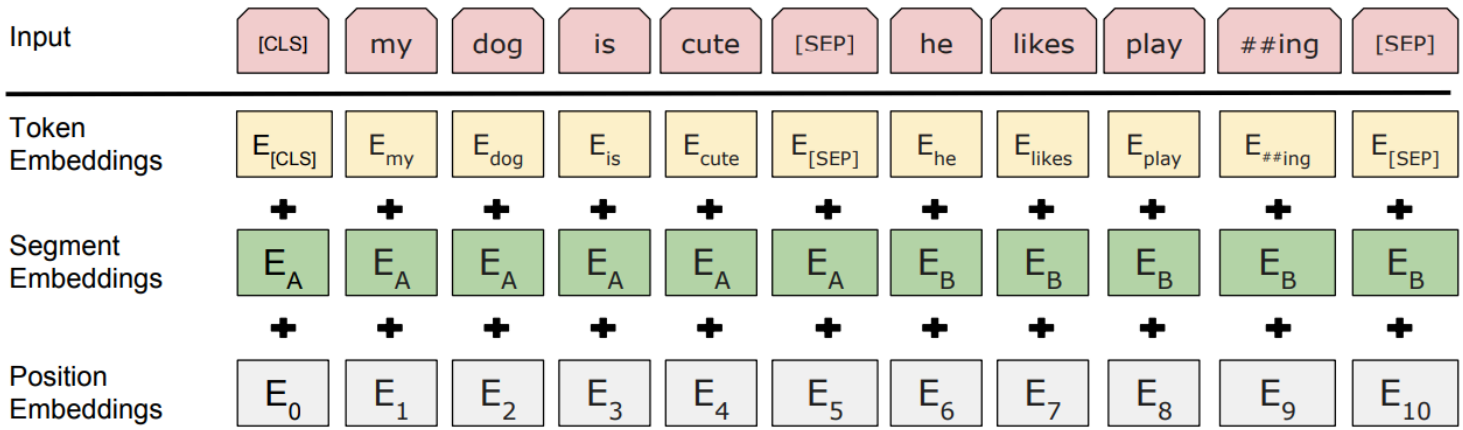
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- **采用了** Transformer **来对文本进行编码**
- **在预训练时采用了多任务学习**（Masked Language Model & Next Sentence Prediction），**与** GPT、ELMo **有显著差别**
- **采用了大量的训练数据**（BooksCorpus & English Wikipedia）
- **采用了大量的模型参数**（12 **层、**24 **层的** Transformer）



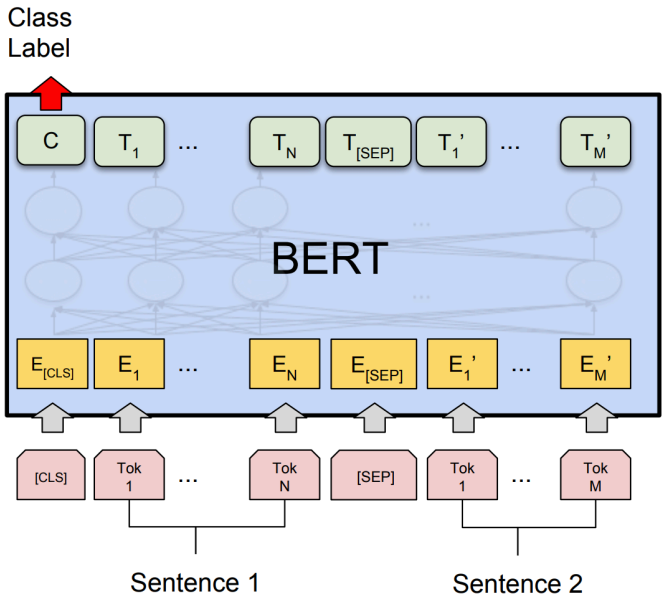
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- **输入上采用了 subword，位置特征，段落特征，能够一定程度上解决 OOV 问题**
- **采用特殊的 [SEP]、[CLS] 来适应预训练任务，捕捉段落特征以及全局特征**

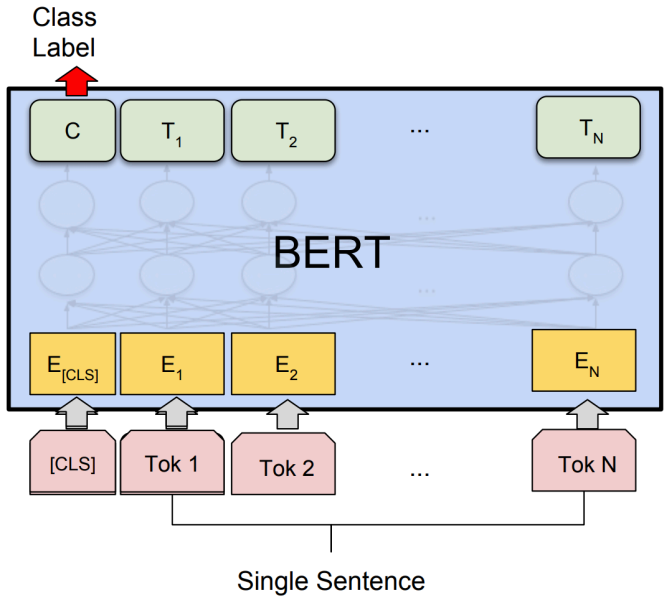


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- **在多数常见 NLP 任务上可以直接微调，且效果显著**



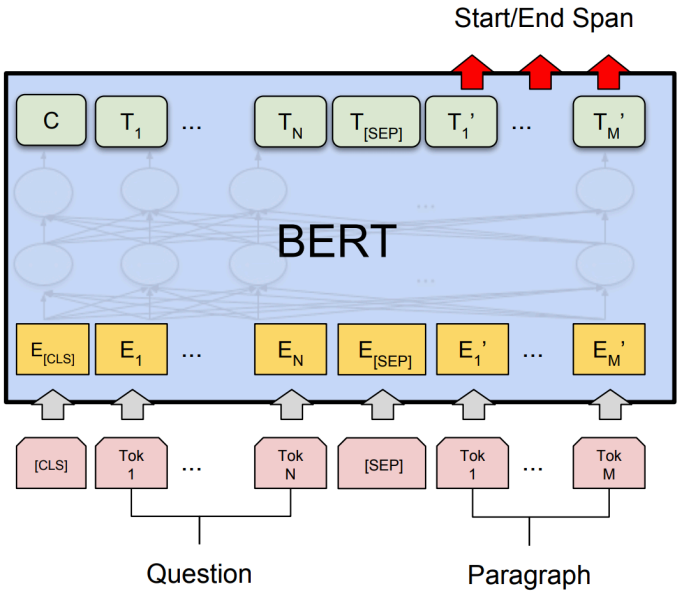
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



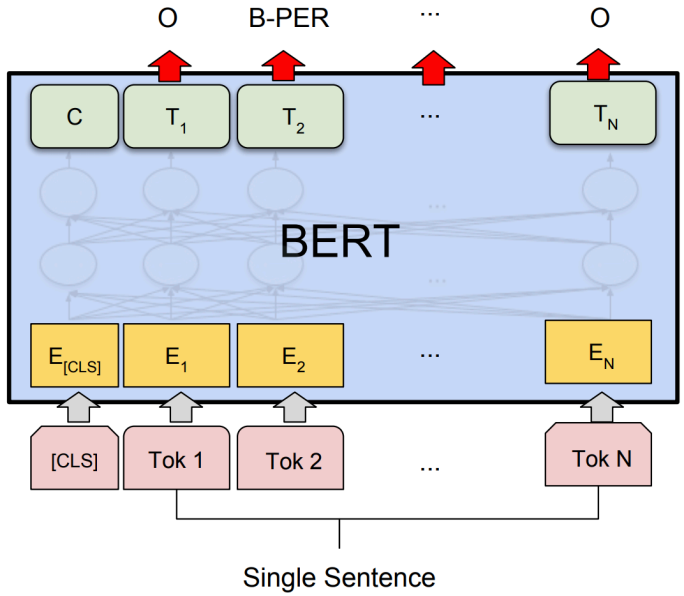
(b) Single Sentence Classification Tasks: SST-2, CoLA

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

- **在多数常见 NLP 任务上可以直接微调，且效果显著**



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

文本图谱联合训练

文本图谱联合训练

- **在 Word2vec 时代，联合文本和图谱学习更好的向量表示**
- Knowledge graph and text jointly embedding. (Wang et al., 2014)
- Representing text for joint embedding of text and knowledge bases. (Toutanova et al., 2014)
- Joint representation learning of text and knowledge for knowledge graph completion. (Han et al., 2016)
- Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. (Yamada et al., 2016)

Knowledge graph and text jointly embedding

- 核心思路：将实体向量和单词向量在同一个向量空间中进行学习
- 定义三种条件概率

$$\Pr(h|r,t) = \frac{\exp\{z(\mathbf{h},\mathbf{r},\mathbf{t})\}}{\sum_{\tilde{h}\in\mathcal{I}} \exp\{z(\tilde{\mathbf{h}},\mathbf{r},\mathbf{t})\}}$$

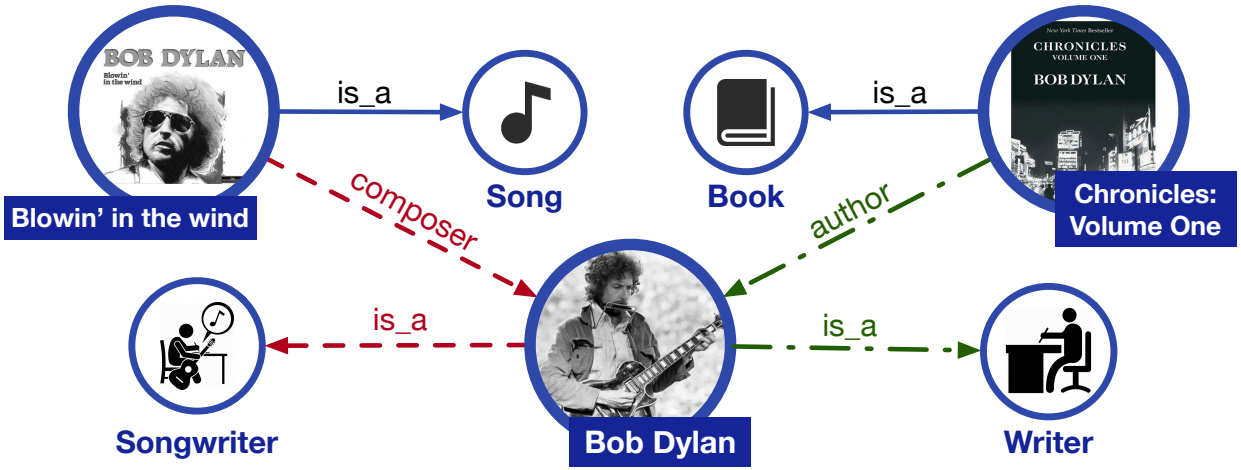
$$\Pr(w|v) = \frac{\exp\{z(\mathbf{w}',\mathbf{v})\}}{\sum_{\tilde{w}\in\mathcal{V}} \exp\{z(\tilde{\mathbf{w}}',\mathbf{v})\}}$$

$$\Pr(w|e) = \frac{\exp\{z(\mathbf{w}',\mathbf{e})\}}{\sum_{\tilde{w}\in\mathcal{V}} \exp\{z(\tilde{\mathbf{w}}',\mathbf{e})\}}$$

工作动机

工作动机

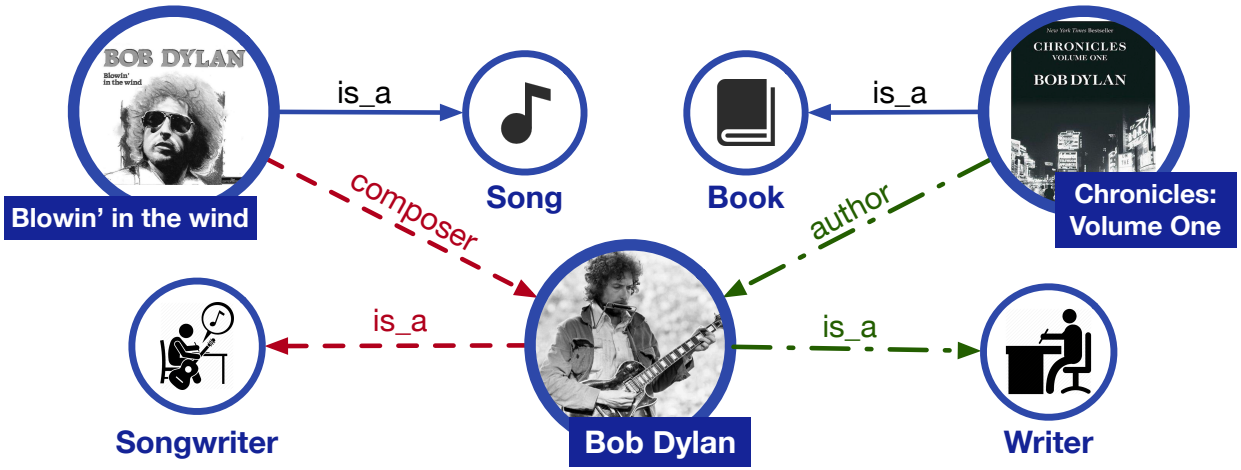
- **现有语言表示模型难以捕捉低频实体信息**
  ELMo: Character CNN
  BERT: sub-word



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

工作动机

- 外部的知识信息可以增强语言表示模型
- 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）

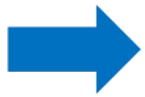


Bob Dylan wrote ***Blowin' in the Wind*** in 1962, and wrote ***Chronicles: Volume One*** in 2004.

模型结构

基于知识增强的语言表示模型

- 知识图谱作为一个重要的外部知识来源，可以提供丰富的知识信息
- ERNIE: 一个在大规模语料和知识图谱上预训练的语言表示模型



两个挑战

- **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）
  - 单词信息
  - 句法信息
  - 知识信息

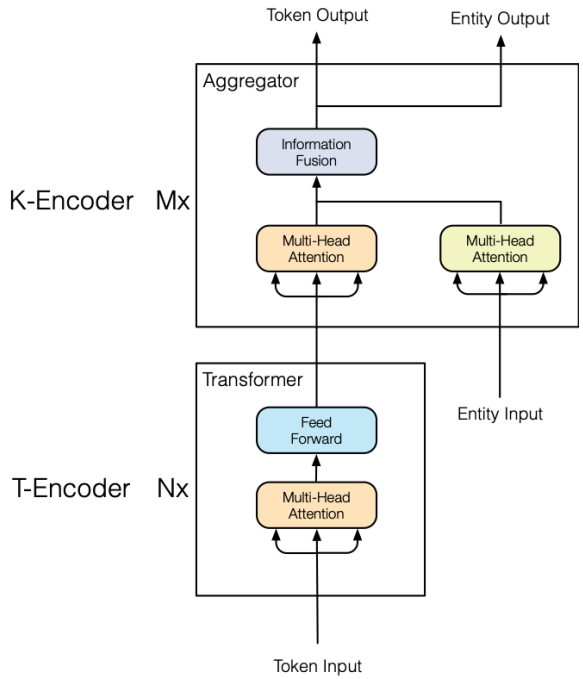
结构化知识表示

- 根据文本检索相关知识
  将命名实体短语链接到图谱当中的实体（TAGME、XLink）
  通过文本中的实体把知识信息引入到预训练模型中

- 表示结构化信息
  使用知识表示算法编码知识图谱的结构信息（TransE）
  将学习到的实体表示作为模型的输入特征

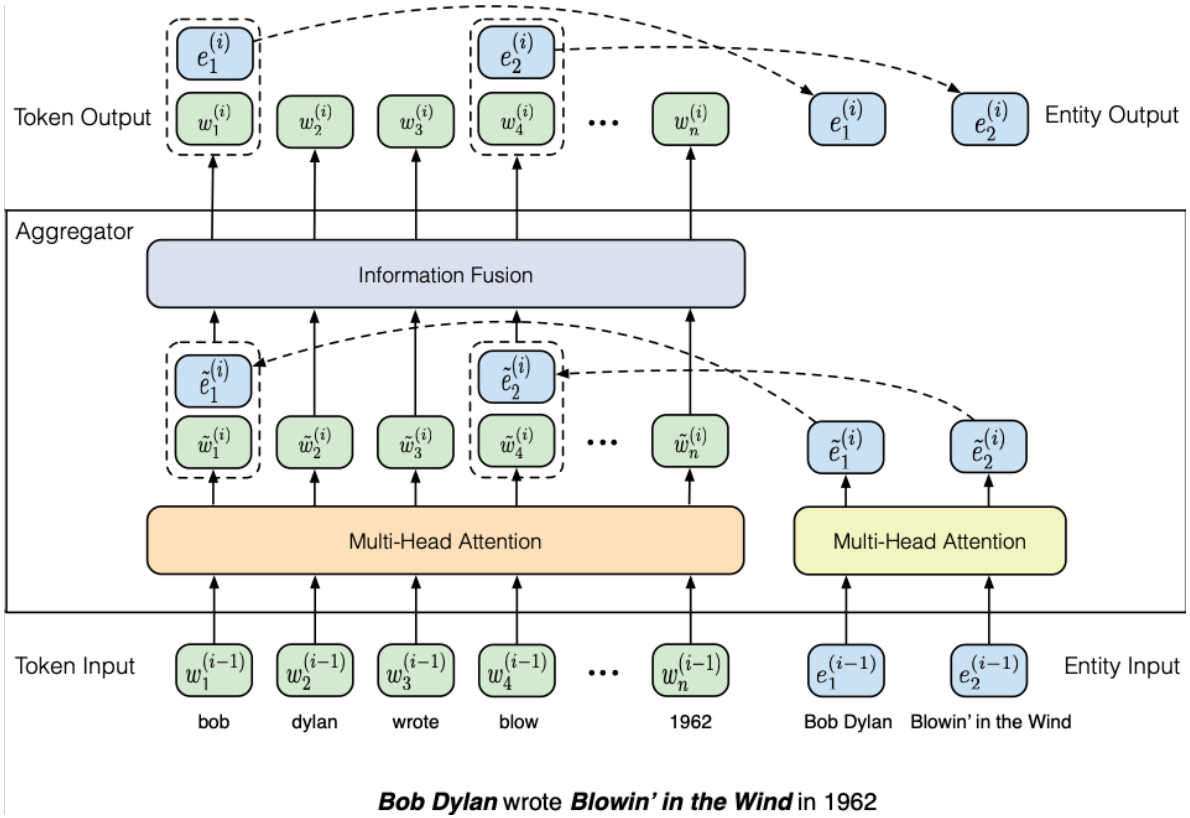
结构化知识表示

- ERNIE 的模型结构
  底层模型对于文本进行建模
  高层模型对于知识信息进行整合



结构化知识表示

- 为了进行知识融合，我们设计了收集器层



异质信息融合

- 收集器层中的信息融合层有两种输入方式
- 有对应实体的单词

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

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

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

异质信息融合

- 收集器层中的信息融合层有两种输入方式
- 无对应实体的单词

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

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

异质信息融合

- **降噪实体自编码机**（Denoising Entity Auto-encoder）
- **使用输出的词向量预测对应的实体**

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

- **加入噪声增强模型的鲁棒性**
  - 5%: **随机替换输入实体**
  - 15%: **遮蔽输入的实体**
  - 80%: **输入正确的实体**

实验部分

预训练细节

- **文本数据：** Wikipedia（45 **亿个单词，**1.4 **亿个实体**）
- **图谱数据：** Wikidata（500 **万个实体，**2000 **万个事实**）
- 6 **层** T-Encoder，6 **层** K-Encoder
- $H_w = 768, H_e = 100, A_w = 12, A_e = 4$
- **参数总量** 114M（BERT **为** 101M）
- **使用** $\text{BERT}_{\text{BASE}}$ **初始化**
- **在语料上训练一轮收敛**

微调细节

- 在普通任务和知识驱动的任务上微调
  普通任务与 BERT 一致
  知识驱动任务微调需要关注于文本中的实体短语

Mark Twain wrote ***The Million Pound Bank Note*** in 1893.

Input for Common NLP tasks:

Input for Entity Typing:

Input for Relation Classification:

关系分类

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 1: 多个模型在 FewRel 和 TACRED 上关系分类实验的结果 (%).

实体分类

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**

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 2: 模型在 FIGER(上) 和 Open Entity(下) 数据集上的结果 (%).

GLUE

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 3: BERT 和 ERNIE 在 GLUE 数据集上不同任务的效果 (%).

总结

总结

- ERNIE 尝试将知识信息引入语言表示模型，在知识驱动任务上取得了优于 BERT 的结果
  - 融合知识的聚合器网络
  - 预训练任务降噪实体自编码机
- 未来方向
  - 将知识注入基于特征的预训练模型，如 ELMo
  - 将各种结构化知识引入语言表示模型，如 ConceptNet
  - 启发式标注更多的现实世界语料库

谢谢！

Paper

Code