KG-BERT: BERT for Knowledge Graph Completion

jtwang 2022/03/20 [1]

> Yao L, Mao C, Luo Y. KG-BERT: BERT for knowledge graph completion[J]. arXiv preprint arXiv:1909.03193, 2019.

Abstract

使用预训练的语言模型补全KG。将三元组作为文本序列，提出了一个新的框架KG-BERT（knowledge graph bidirectional encoder representations from transformer）。将一个三元组中实体和关系的描述作为输入，使用KG-BERT语言模型计算三元组的评分函数。达到了SOTA效果。

Introduction

Motivation：大多数KGE方法只利用structure information，suffer from sparseness。最近的研究融合了textual information，但是对于不同三元组中相同的实体/关系只学习一个文本嵌入，忽略了contextual information。例如，different words in the description of Steve Jobs should have distinct importance weights connected to two relations “founded” and “isCitizenOf”【？？】；关系“wroteMusicFor“在不同的实体下有两种不同的含义”write lyrics“和”composes musical compositions“。另一方面大规模语料库中的syntactic and semantic信息没有被充分利用，因为现有的方法只采用了实体描述、relation mentions【？？是什么】、word co-occurrence with entities。

最近，预训练的语言模型例如ELMo、GPT、BERT、XLNet在自然语言处理领域取得了巨大成功。这些模型使用大量free text data学习contextualized word embeddings，在许多自然语言理解任务中取得了SOTA效果。其中，BERT是最prominent，通过masked language modeling和next sentence prediction两个任务，预训练双向Transformer encoder，可以capture rich linguistic knowledge。

本文我们使用预训练语言模型用于知识图谱补全。我们将实体、关系、三元组看作文本序列，将KG补全看作序列分类问题。通过微调BERT，预测一个三元组的plausibility。这个方法在许多KG补全任务上achieve strong performance。本文的贡献

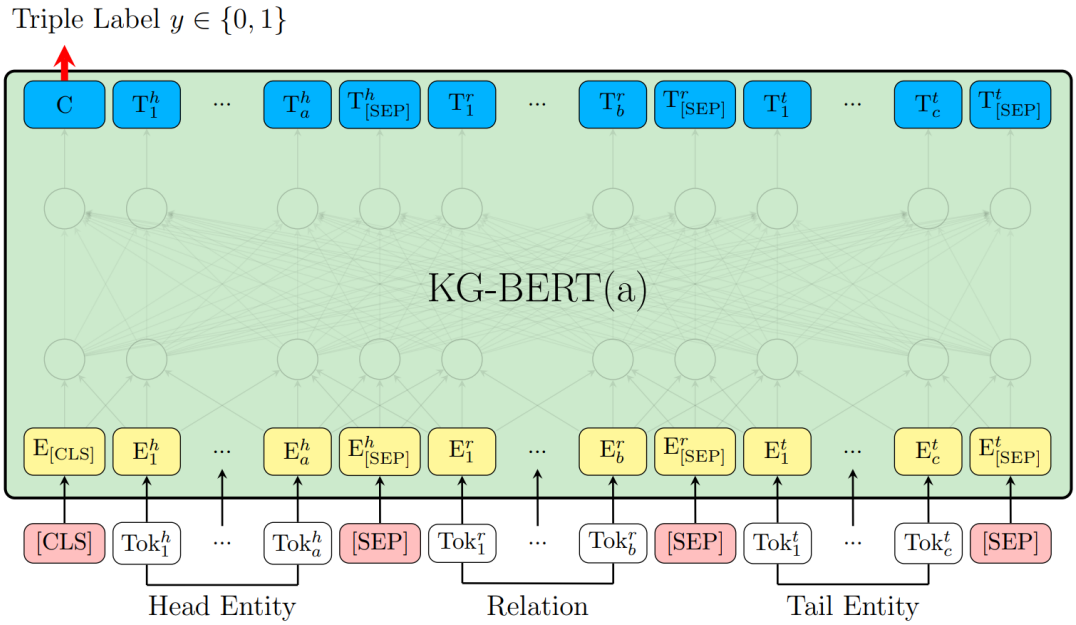
- 提出一个新的language modeling method for KG completion，第一个使用预训练语言模型建模三元组的plausibility
- 结果表明我们的方法在三元组分类、关系预测和连接预测任务中取得了SOTA效果

Method

BERT：基于双向Transformer编码器，是SOTA预训练contextual语言表示模型。Transformer编码器基于self-attention机制。BERT框架分为两步：pre-training和fine-tuning。在预训练时，BERT在大型无标注general domain语料库上，使用两个自监督任务masked language modeling和next sentence prediction进行训练。在微调时，BERT先初始化为预训练的参数权重，然后使用下游任务（sentence pair classification、question answering、sequence labeling）的标注数据进行微调。

KG-BERT：微调预训练的BERT用于KG补全。使用名称或描述（name/description）表示实体和关系，使用name/description单词序列作为BERT的input sentence。（As original BERT, a “sentence” can be an arbitrary span of contiguous text or word sequence, rather than an actual linguistic sentence.）我们pack the sentences of (h, r, t) 作为一个单独的序列（sequence），一个sequence对应BERT中的input token sequence。

KG-BERT(a) for triple classification



输入序列：每个输入序列都以一个特殊的token [CLS]开头，(h, r, t)之间使用[SEP]分隔开

- 头实体被表示为一个包含许多token $Tok_1^h, \ldots, Tok_a^h$的语句，例如：“Steven Paul Jobs was an American business magnate, entrepreneur and investor.”【description】或者”Steven Jobs“【name】
- 关系表示为包含$Tok_1^r, \ldots, Tok_b^r$的语句，例如：”founded“
- 尾实体表示为包含$Tok_1^t, \ldots, Tok_c^t$的语句，例如：“Apple Inc. is an American multinational technology company headquartered in Cupertino, California.” or “Apple Inc.”

输入表示：对于一个token，它的input representation由token+segment+position embeddings构成。每个input token i 有一个input representation E_i。

- 头尾实体tokens共享相同的segment embedding e_A，关系tokens的segment embedding为 e_B 。
- 在相同位置上 $i \in 1, 2, \ldots, 512$ 的不同tokens有相同的position embedding

Token representations【$E_{[CLS]}, E_i$】被送入BERT模型，[CLS]的最后一层隐藏向量表示为$C \in \mathbb{R}^H$，E_i的最后一层隐藏向量表示为$T_i \in \mathbb{R}^H$，H是隐层状态size。C被用来计算三元组得分。

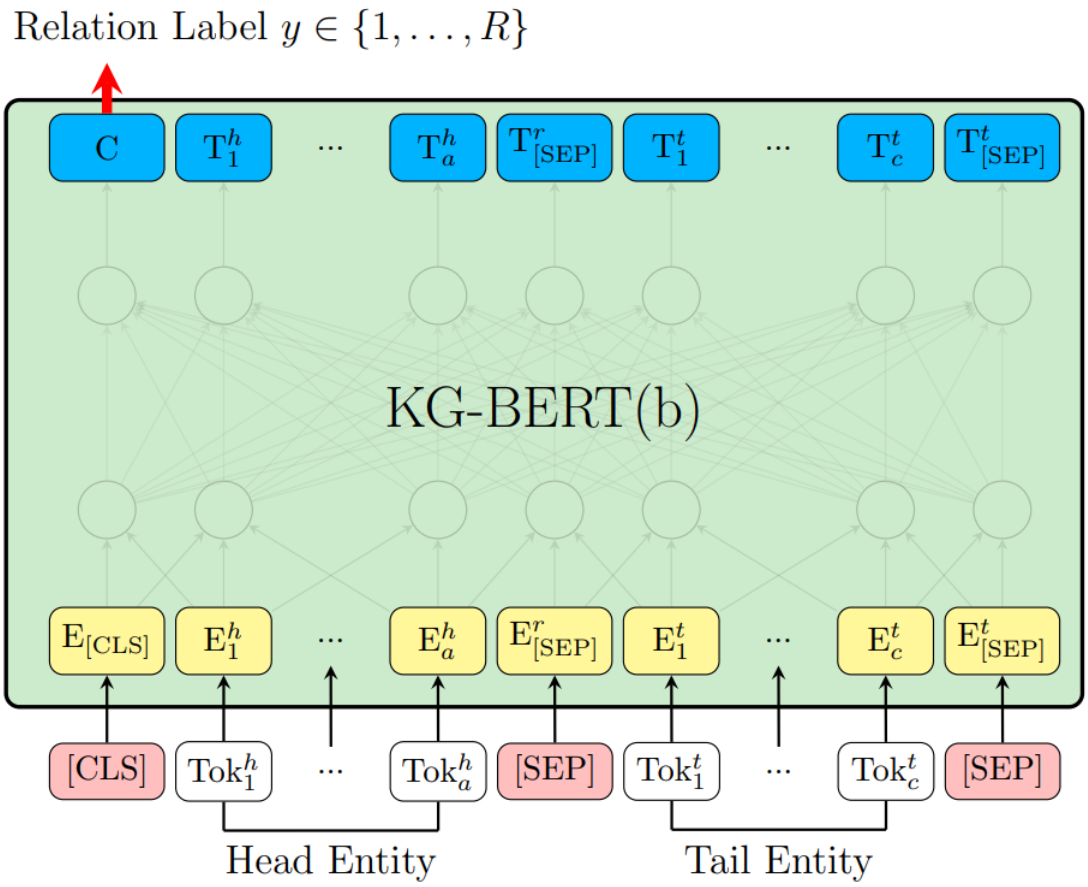
唯一新引入的参数是分类层权重$W \in \mathbb{R}^{2 \times H}$，一个三元组的评分函数如下，结果是一个2维向量，表示分类结果为0/1的概率。

$$s_\tau = f(h, r, t) = sigmoid(CW^T) \in \mathbb{R}^2$$

给定正三元组集合$\mathbb{D}^+$，负三元组集合$\mathbb{D}^-$，使用交叉熵损失，其中$y_\tau \in 0, 1$是标签。

$$\mathcal{L} = -\sum_{\tau \in \mathbb{D}^+ \cup \mathbb{D}^-} (y_\tau log(s_{\tau 0}) + (1 - y_\tau) log(s_{\tau 1}))$$

KG-BERT(b) for predicting relations



只使用两个实体h和t预测它们之间的关系r。这样做比KG-BERT(a) with relation corruption的效果更好。与KG-BERT(a)相同，C被用来表示两个实体。新引入参数$W \in \mathbb{R}^{R \times H}$，其中R是KG中关系的数量。三元组的评分函数为

$$s'_\tau = f(h, r, t) = softmax(CW'^T) \in \mathbb{R}^R$$

使用交叉熵损失

$$\mathcal{L}' = -\sum_{\tau \in \mathbb{D}^+} \sum_{i=1}^{R} y'_{\tau i} log(s'_{\tau i})$$

Experiments

数据集：WN11、FB13、FB15K、WN18RR、 FB15k-237、UMLS

Dataset	# Ent	# Rel	# Train	# Dev	# Test
WN11	38,696	11	112,581	2,609	10,544
FB13	75,043	13	316,232	5,908	23,733
WN18RR	40,943	11	86,835	3,034	3,134
FB15K	14,951	1,345	483,142	50,000	59,071
FB15k-237	14,541	237	272,115	17,535	20,466
UMLS	135	46	5,216	652	661

Settings：使用pre-trained BERT-Base with 12 layers, 12 self-attention heads and H = 768初始化KG-BERT，使用Adam微调。实验发现BERT-Base比BERT-Large表现更好，而且BERT-Base更简单，对超参选择less sensitive。

Triple Classification

Method	WN11	FB13	Avg.
NTN (Socher et al. 2013)	86.2	90.0	88.1
TransE (Wang et al. 2014b)	75.9	81.5	78.7
TransH (Wang et al. 2014b)	78.8	83.3	81.1
TransR (Lin et al. 2015b)	85.9	82.5	84.2
TransD (Ji et al. 2015)	86.4	89.1	87.8
TEKE (Wang and Li 2016)	86.1	84.2	85.2
TransG (Xiao, Huang, and Zhu 2016)	87.4	87.3	87.4
TranSparse-S (Ji et al. 2016)	86.4	88.2	87.3
DistMult (Zhang et al. 2018)	87.1	86.2	86.7
DistMult-HRS (Zhang et al. 2018)	88.9	89.0	89.0
AATE (An et al. 2018)	88.0	87.2	87.6
ConvKB (Nguyen et al. 2018a)	87.6	88.8	88.2
DOLORES (Wang, Kulkarni, and Wang 2018)	87.5	89.3	88.4
KG-BERT(a)	**93.5**	**90.4**	**91.9**

KG-BERT表现好的主要原因

- 输入序列包含both entity and relation word sequences
- 三元组分类任务与next sentence prediction task很相似
- 相同的token在不同的三元组中会有不同的隐藏向量，利用了contextual information
- self-attention机制可以发掘与三元组事实相关的最重要的单词

Link Prediction

Method	WN18RR		FB15k-237		UMLS	
	MR	Hits@10	MR	Hits@10	MR	Hits@10
TransE (our results)	2365	50.5	223	47.4	1.84	98.9
TransH (our results)	2524	50.3	255	48.6	1.80	**99.5**
TransR (our results)	3166	50.7	237	51.1	1.81	99.4
TransD (our results)	2768	50.7	246	48.4	1.71	99.3
DistMult (our results)	3704	47.7	411	41.9	5.52	84.6
ComplEx (our results)	3921	48.3	508	43.4	2.59	96.7
ConvE (Dettmers et al. 2018)	5277	48	246	49.1	–	–
ConvKB (Nguyen et al. 2018a)	2554	52.5	257	51.7	–	–
R-GCN (Schlichtkrull et al. 2018)	–	–	–	41.7	–	–
KBGAN (Cai and Wang 2018)	–	48.1	–	45.8	–	–
RotatE (Sun et al. 2019)	3340	**57.1**	177	**53.3**	–	–
KG-BERT(a)	**97**	52.4	**153**	42.0	**1.47**	99.0

Relation Prediction

Method	Mean Rank	Hits@1
TransE (Lin et al. 2015a)	2.5	84.3
TransR (Xie, Liu, and Sun 2016)	2.1	91.6
DKRL (CNN) (Xie et al. 2016)	2.5	89.0
DKRL (CNN) + TransE (Xie et al. 2016)	2.0	90.8
DKRL (CBOW) (Xie et al. 2016)	2.5	82.7
TKRL (RHE) (Xie, Liu, and Sun 2016)	1.7	92.8
TKRL (RHE) (Xie, Liu, and Sun 2016)	1.8	92.5
PTransE (ADD, len-2 path) (Lin et al. 2015a)	**1.2**	93.6
PTransE (RNN, len-2 path) (Lin et al. 2015a)	1.4	93.2
PTransE (ADD, len-3 path) (Lin et al. 2015a)	1.4	94.0
SSP (Xiao et al. 2017)	**1.2**	–
ProjE (pointwise) (Shi and Weninger 2017)	1.3	95.6
ProjE (listwise) (Shi and Weninger 2017)	**1.2**	95.7
ProjE (wlistwise) (Shi and Weninger 2017)	**1.2**	95.6
KG-BERT (b)	**1.2**	**96.0**

Attention Visualization

以WN18RR中一个正三元组(__twenty_dollar_bill_NN_1, _hypernym, _note_NN_6)为例，实体描述是"a United States bill worth 20 dollars"和"a piece of paper money"，关系名称是"hypernym"。我们发现一些中重要的单词如"paper"和"money"有较高的注意力得分，一些不相关的单词如"united"和"states"分数则较低。

不同的attention head关注不同的tokens。有6个attention heads关注[SEP]，3个关注"a" and "piece"，其他4个heads关注"paper"和"money"。multi-head attention allows KG-BERT to jointly attend to information from different representation subspaces at different positions, different attention heads are concatenated to compute the final attention values。【multi-head attention的好处】

Discussions：BERT模型的一个主要缺点是expensive，这使得link predication的评估非常耗时。一个可能的解决方法是使用ConvE中的1-N scoring，或者使用lightweight语言模型。

- 图片地址

1. 【】里是我的注释