文章编号：1003-0077（2011）00-0000-00

基于带注意力机制 CNN 的联合知识表示模型*

彭敏，姚亚兰，谢倩倩，高望

（武汉大学 计算机学院，武汉 430072）

摘要：知识表示学习在自然语言处理领域获得了广泛关注，尤其在实体链指、关系抽取及自动问答等任务上表现优异。然而，大部分已有的表示学习模型仅利用知识库中的结构信息，无法很好地处理新的实体或关联事实极少的实体。为解决该问题，本文提出了引入实体描述信息的联合知识表示模型。该模型先利用卷积神经网络编码实体描述，然后利用注意力机制来选择文本中的有效信息，接着又引入位置向量作为补充信息，最后利用门机制联合结构和文本的向量，形成最终的联合表示。实验表明，本文的模型在链路预测和三元组分类任务上与目前最好的模型性能相近。

关键词：知识表示学习；卷积神经网络；注意力机制

中图分类号：TP391　　**文献标识码：**A

Knowledge representation learning for joint structural and textual embed-ding via Attention-based CNN

Min Peng, Yalan Yao, Qianqian Xie, Wang Gao

(School of Computer, Wuhan University, Wuhan 430072, China)

Abstract: Knowledge representation learning has attracted lots of attention in natural language processing with especially encouraging results on tasks such as Entity Linking, Relationship Extraction, Question Answering and so on. However, most of the existing models only use the structural information of knowledge graph and cannot handle new entities or entities with few facts well. Hence, we propose a joint knowledge representation model which utilizes both entity description and structural information. Firstly, we introduce convolutional neural network models to encode the entity description. Then, we design the attention mechanism to select the valid information of the text. Moreover, we introduce the position vector as the supplementary information. Finally, a gating mechanism is used to integrate the structural and textual information into the joint representation. Experimental results show that our models outperform other baselines on link prediction and triplet classification tasks.

Key words: knowledge representation learning; CNN; attention mechanism

1 引言

目前，知识库在智能问答及个性推荐等人工智能领域应用前景广泛。知识库通常表示成网络结构，使用三元组（头实体，关系，尾实体）来表示知识。然而，基于网络的知识表示面临以下挑战：（1）计算效率低。知识推理时往往要设计专门的图算法，计算复杂度高且拓展性差。（2）严重的数据稀疏。知识库存在一些关联知识较少的罕见实体，其语义计算准确率极低。基于以上挑战，人们提出了以深度学习为基础的知识表示学习方法，TransE[1]便是其中应用最为广泛的模型。然而，TransE 及其大部分拓展模型仅利用知识库的结构化信息，很难处理好知识库外的新实体或相关知识极少的罕见实体。

为解决以上问题，一些工作[2-4]开始引入文本信息来改善知识表示。面对新实体或罕见实体，它们利用文本来补充其缺失的语义信息，不仅提供了新的表示方法，还能有效缓解数据稀疏。然而，它们仍存在一些不足：（1）尚未提出联合文本和结构化信息的有效方法。许

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61772382）；国家自然科学基金（61472291）

作者简介：彭敏（1973——），女，教授，自然语言处理；姚亚兰（1995——），女，硕士，自然语言处理；谢倩倩（1994——），女，博士，自然语言处理；高望（1983——），男，博士，自然语言处理。

多工作只在单词的层次或得分函数上做了对齐。（2）未筛选文本信息。例如实体描述可能包含实体在多种情况下的信息，并非所有的文本都有用。

针对已有工作的不足，Xu 等人提出基于双向 LSTM 的联合知识表示模型[5]。该模型利用注意力机制来筛选描述文本里的信息，提出了门机制来联合文本和结构表示，在链路预测及三元组分类等经典任务里达到了目前最好的水平。然而，双向 LSTM 模型需要输入上一个隐藏状态和位置来产生下一个隐藏状态，这种固有顺序的性质使得训练过程无法并行化，在处理更长的序列时，还会因内存限制制约训练集的跨批次处理[6]。

基于以上问题及面临的挑战，为联合文本信息来缓解知识库稀疏，准确捕捉文本中最相关的语义，同时考虑到卷积核可并行化及高效计算的优势，本文提出了基于带注意力机制 CNN 的联合知识表示模型 JA-CNN。首先，本文提出了基于 CNN 的文本编码器，然后设计了专门的注意力机制来选择描述文本中与情景最相关的语义信息；其次，本文采用 TransE 模型来编码知识库的结构化信息；最后，本文引入门机制来控制多源信息传递到联合表示的权重，形成最终的表示。此外，本文还提出了基于 JA-CNN 的改进模型 JPA-CNN。该模型尝试在输入端引入位置向量，使编码器也具备捕捉句子词位置信息的能力。在链路预测及三元组分类任务的实验表明，本文的模型能显著改善稀疏问题，各项指标与最先进的方法相比都有很强的竞争性，尤其在关系分类任务下有明显优势。

本文的主要贡献：

（1）本文提出了联合实体描述和结构化信息的联合知识表示模型 JA-CNN。该模型设计专门的注意力机制来捕捉描述文本中的最相关信息，帮助提高实体表示的区分度。

（2）本文提出了基于 JA-CNN 的拓展模型 JPA-CNN。该模型引入位置向量，使 CNN 能捕捉描述句子中的位置信息。

（3）实验结果表明，本文的模型达到与目前最好的模型性能相近，还拥有可并行化及高效计算的优势。

2 相关工作

2.1 知识表示学习

近年来，知识表示学习在知识获取、融合及推理等多种任务里均表现优异，一度成为研究热点。

Bordes 等人提出非结构化模型（Unstructured model）[7]，该模型假设头尾实体向量相似，在得分函数里将关系向量设置为零，因此无法区分不同关系。Bordes 等人提出结构化模型（structured embedding, SE）[8]，该模型假定头尾实体向量只在相关关系的语义空间内相似。此外，Bordes 等人提出语义匹配能量模型（semantic matching energy, SME）[9]，利用投影矩阵表示实体与关系，根据得分函数分为线性形式（linear）和双线性形式（Bilinear）。之后，Bordes 等人提出 TransE 模型[1]，该模型简单高效易拓展，逐渐成为最受关注的知识表示模型。

TransE 将关系表示为从头实体到尾实体的平移向量，旨在将知识库中的实体和关系投影到同一个低维向量空间。Wang 等人提出 TransH 模型[10]，将关系建模为超平面并将头尾实体投影到关系特定的超平面，解决了 TransE 的实体在不同关系下无法有不同表示的问题。Lin 等人提出 TransR 模型[11]，在不同语义空间内表示实体和关系，并将实体投影到对应的关系空间。Lin 等人进一步提出了 CTransR 模型[11]，利用聚类划分关系，为每个关系分别学习表示向量。Ji 等人提出 TransD 模型[12]，利用投影矩阵将头实体和尾实体分别投影到关系空间，解决了 TransR 参数过多的问题。Ji 等人提出 TranSparse 模型[13]，将 TransR 模型中的稠密矩阵换成稀疏矩阵，头、尾实体都有投影矩阵，其中矩阵的稀疏度由关系连接实体的数量决定。Xiao 等人提出 TransA 模型[14]，使用马氏距离替换得分函数中的距离。He 等人提出 KG2E 模型[15]，利用高斯分布来表示实体及关系。Xiao 等人提出 TransG 模型[16]，使用高

斯混合模型表示关系，使关系能包含多种语义。

这些工作仅利用知识库的结构信息，未能有效利用与知识库相关的其它信息如实体描述等。考虑到多源信息能缓解数据稀疏，提高知识表示的区分度，研究者们开始尝试融合多源信息来改善知识表示。

2.2 引入文本信息的知识表示

目前，已有许多研究工作使用文本信息来改善知识表示。

Socher 等人提出 NTN 模型[17]，使用实体名称的词向量平均值来表示实体。Wang 等人通过对齐实体名称和维基百科锚点，将知识和文本投影到同一空间，提高了事实预测的准确性[2]。Zhong 等人在 Wang 等人工作的基础上拓展模型[3]，将实体描述中的知识和词汇关联起来。然而，这两份工作都在词级别上做了对齐，导致其丢失短语或句子层面的语义信息。Zhang 等人使用实体名称或者实体描述中词向量的平均值[18]，该方法忽略了句子中的词序信息。

Xie 等人提出了 DKRL 模型[4]，利用实体描述来表示实体向量。该模型使用连续词袋模型和卷积神经网络来编码实体描述的语义，并将得分函数分成基于结构和基于描述的两部分。尽管该模型也使用 CNN 编码文本信息，但它的 CNN 只包括卷积层、非线性层和池化层，与本文的 CNN 结构有一定差别。此外，该方法尚未考虑到文本信息的筛选及联合两种表示的有效方式。Xu 等人提出了基于双向 LSTM 的联合表示模型[5]，利用注意力机制选择实体描述中的相关文本，同时设计门机制来控制结构信息和文本信息的权重。该方法相比先前的模型性能显著提高，但双向 LSTM 模型的隐状态需要按序生成，训练时无法并行处理，制约了面对长序列的计算效率。

除了实体描述外，还有一些工作[19-21]将文本关系和知识库关系映射到相同的向量空间并获得显著改进。

3 模型介绍

本文的联合模型主要分为三部分：基于 TransE 的结构表示、基于 CNN/A-CNN/PA-CNN 的文本表示和基于门机制的多源信息融合。首先，本文利用 TransE 来编码三元组的结构信息；然后设计了三种编码实体描述的文本编码器：CNN、引入注意力机制的 A-CNN 和在 A-CNN 的基础上引入位置信息的 PA-CNN；最后利用门机制决定结构表示和文本表示构成联合表示的权重。图 1 展示了联合知识表示的整体框架。下面将对模型的每个层次的功能进行详细阐述。

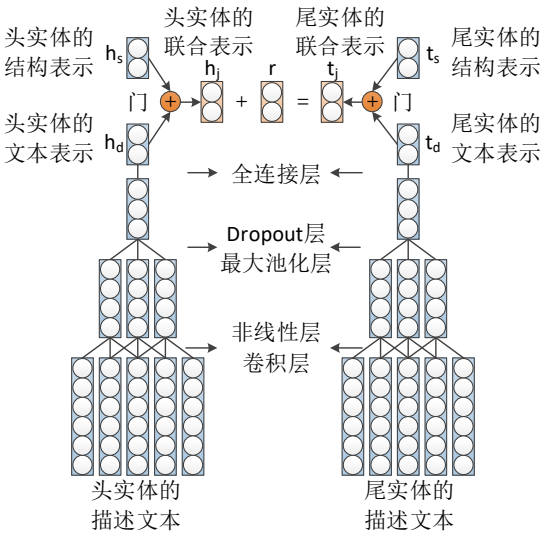

图 1 联合知识表示的整体框架

3.1 基于 TransE 的结构表示

基于 TransE 的表示模型在知识推理、关系抽取等任务里表现优异，也成为知识表示的研究热点。

给定三元组（头实体，关系，尾实体），将其表示为 (h,r,t)。三元组 (h,r,t) 对应的向量表示为 $(\mathbf{h},\mathbf{r},\mathbf{t})$。TransE 旨在将实体和关系表示成低维连续的向量。合法的三元组的向量应该满足以下公式：$\mathbf{h}+\mathbf{r}\approx\mathbf{t}$，错误的三元组则不满足。因此，TransE 定义了如下得分函数来衡量三元组的质量：

$$f_r(h,t)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{L_1/L_2},\ \text{s.t.}:\|\mathbf{h}\|_2^2\le 1;\ \|\mathbf{t}\|_2^2\le 1 \tag{1}$$

式(1)即向量 $\mathbf{h}+\mathbf{r}$ 和 $\mathbf{t}$ 的 L_1 或 L_2 距离。对于合理的得分函数，合法三元组的得分要比错误三元组的得分更低。

3.2 基于 CNN/A-CNN/PA-CNN 的文本表示

目前，大型知识库中的实体通常都有其对应的实体描述信息。实体描述包含实体在各种情景下的语义信息，有助于改善实体表示，使其区分度更强，同时也能缓解数据稀疏。

本文需从不定长的实体描述中编码文本信息。考虑到卷积核能捕捉文本信息的局部特征，拥有可并行化，运行速度快等优点，本文最终选择基于 CNN 的文本编码方式。

3.2.1 基于 CNN 的文本表示

文本预处理：本文先去除实体描述里的标点符号，然后使用 Word2Vec 预先训练好的词向量[22]初始化词序列，以此作为 CNN 的输入。

卷积层：卷积层的输入是预处理后长度为 n 的词序列 x，本文定义为 $x_{1:n}=x_1,x_2,\mathrm{K},x_n$，其中 $x_i\in\mathbb{R}^d$ 表示句子中第 i 个词语的 d 维词向量。

对词序列 x，卷积层选取大小为 k 的滑动窗口内的词序列进行卷积操作，输出特征映射 c。词序列的长度不固定，本文以词序列的最大长度 n 为标准，在所有长度不符合的词序列末尾填充零向量，得到定长输入。

滑动窗口处理的词序列定义为：

$$x_{i:i+k-1}=x_i,x_{i+1},\mathrm{K},x_{i+k-1} \tag{2}$$

窗口内词序列卷积后输出的第 i 个向量为：

$$c_i=f(w\cdot x_{i:i+k-1}+b) \tag{3}$$

其中，$w\in\mathbb{R}^{k\times d}$ 是滤波器，$b\in\mathbb{R}$ 是偏置项，f 是激活函数，本文选取线性整流函数 ReLU 作为激活函数。

卷积的边界处理（padding）设置为 SAME，即使用零填充。卷积层的输出为：

$$c=[c_1,\ldots,c_n] \tag{4}$$

池化层：本文采用最大池化，对每个窗口内的输入向量选取最大值构成新向量。

窗口大小为 n_p 的池化层输出的第 i 个向量为：

$$p_i=\max(c_{n_p\cdot i},\mathrm{K},c_{n_p\cdot(i+1)-1}) \tag{5}$$

滤波器的数量为 m，池化层的输出为 $p=[p_1,\ldots,p_m]$。

Dropout 层：在训练期间，本文通过伯努利函数得到向量 r 与输入进行元素相乘，用于屏蔽部分隐层神经元，防止过拟合。

Dropout 层的输出定义为：

$$r:\ \text{Bernoulli}(\rho),\ y=r*p \tag{6}$$

其中，Bernoulli 函数是以概率 ρ 随机生成 0 或 1 的向量，用于移除神经元。

全连接层：对输入进行矩阵向量乘积操作得到网络的最终输出向量。

CNN 的输出定义为：

$$e_d = w_o \cdot [\cdot] + b_o \tag{7}$$

其中，w_o 是参数矩阵，b_o 是可选偏置。

3.2.2 基于 CNN 的文本表示

CNN 对整体描述文本进行语义编码，没有考虑到描述信息包含实体在多种关系下的不同语义。这意味着给定三元组后关系特定，描述里包含的其它关系的信息会造成一定干扰。因此，本文基于 CNN 提出文本编码器 A-CNN，设计了相应的注意力机制，通过三元组的关系来捕捉描述中与其最相关的信息。

对实体描述的词序列 $x_{1:n} = x_1, x_2, \mathrm{K}, x_n$，给定关系 $\mathbf{r} \in \mathbb{R}^d$，$\mathbf{r}$ 拓展一维后得到矩阵 $\hat{\mathbf{r}} \in \mathbb{R}^{d \times 1}$，该描述的注意力定义为：

$$\alpha(\mathbf{r}) = \mathrm{Softmax}(x_{1:n} \hat{\mathbf{r}}) \tag{8}$$

卷积层的输出为 c，添加注意力权重后形成输出 [·]，[·]接下来会作为池化层的输入。[·]的定义如下：

$$[\cdot] = c\alpha(\mathbf{r}) \tag{9}$$

3.2.3 基于 PA-CNN 的文本表示

考虑到 CNN 编码文本时未包括词的顺序特征，可能会丢失部分语义，本文引入词的位置编码作为补充信息。本文基于 A-CNN 提出了文本编码器 PA-CNN，采用 Sukhbaatar 等人提出的方法[23]来编码位置信息。输入向量 I 的第 j 个分量 I_j 由位置向量的分量 l_j 和词向量的分量 x_j 构成。

位置向量 l_j 是一个列向量，拥有以下结构：

$$l_{kj} = (1 - \mathrm{j}/J) - (k/d)(1 - 2j/J) \tag{10}$$

其中，J 是句子中词的个数，d 是位置向量的维度，k 是 l_j 的第 k 个分量。这里位置编码采用和词向量同样的维度，方便将两者相加。

给定长度为 n 的词序列 $x_{1:n} = (x_1, \mathrm{K}, x_n)$，其位置向量为 $l_{1:n} = (l_1, \mathrm{K}, l_n)$，加入位置信息后编码器的新输入为 $I_{1:n} = (x_1 + l_1, \mathrm{K}, x_n + l_n)$。

3.3 基于门机制的多源信息融合

结构信息和文本描述都提供了实体的有效信息，本文采用 Xu 等人提出的门机制[5]将两种信息源整合成联合表示，即将联合表示 e_j 当作结构表示 e_s 和文本表示 e_d 加权求和的结果。

联合表示 e_j 定义为：

$$e_j = g_s \mathrm{e}\ e_s + g_d \mathrm{e}\ e_d,\ \mathrm{s.t.}: g_d = 1 - g_s;\ g_d, g_s \in [0,1] \tag{11}$$

其中，g_s 和 g_d 是平衡两种信息源的门，e 是元素乘法。

门 g 定义为：

$$g = \mathrm{Softmax}([\cdot]) \tag{12}$$

其中，[·]~ uniform(0,1) 是存储在查找表中的实值向量，服从均匀分布。本文利用 Softmax 函数将门的值约束到 [0,1] 之间。

类似 TransE，联合表示的得分函数定义为：

$$f_r(h,t;d_h,d_t) = \| (g_{hs} \mathrm{e}\ h_s + g_{hd} \mathrm{e}\ h_d) + r - (g_{ts} \mathrm{e}\ t_s + g_{td} \mathrm{e}\ t_d) \|_2^2 \tag{13}$$

其中，g_{hs}, g_{hd} 分别是头实体的门，g_{ts}, g_{td} 分别是尾实体的门。

3.4 训练

与 TransE 相似，本文也采用最大间隔方法[1]训练模型。本文使用得分函数 $f_r(h,t)$ 来评估三元组的质量。合法三元组拥有较低得分，错误三元组拥有较高得分，则对应的优化目标函数如下：

$$L = \sum_{(h,r,t)\in S_l} \sum_{(h',r',t')\in \bar{S}_l} \max(0, f_r(h,t) + \gamma - f_r(h',t')) \tag{14}$$

其中，S_l 是合法三元组集合，$\bar{S}_l$ 是错误三元组集合，$\gamma > 0$ 是正负样本之间的间距。本文采用随机梯度下降来优化目标函数。

知识库里的三元组都是正样本，负样本需要自行生成。本文采用 Wang 等人提出的方法[10]，设置不同的概率来替换头实体或尾实体。该方法将关系按照两端连接实体的数目分为 1-1、1-N、N-1 和 N-N 四种，如果是 1-N 关系则增大替换头实体的机会，如果是 N-1 关系则增大替换尾实体的机会。该方法能降低产生错误负样本的概率。

对每个三元组，正样本用 $S_l = \{(h_i, r_i, t_i) \mid y_i = 1\}$ 表示，负样本由 $\bar{S}_l = \{(h_i', r_i', t_i') \mid y_i = -1\}$ 表示。

训练集中的错误三元组由式(15)产生：

$$\begin{aligned} \bar{S}_l = \{(h_{neg}, r_k, t_k) \mid h_{neg} \neq h_k \wedge y_k = -1\} \cup \\ \{(h_k, r_k, t_{neg}) \mid t_{neg} \neq t_k \wedge y_k = -1\} \end{aligned} \tag{15}$$

4 实验结果及分析

本文在链路预测和三元组分类两个常规任务上评估模型的性能。

4.1 实验数据

本文使用两个最常用的数据集，分别是语言知识库 WordNet[24]的子集 WN18 和世界知识库 Freebase[25]的子集 FB15k。表 1 展示了数据集的相关属性。

表 1 数据集的统计结果

数据集	关系	实体	训练集	验证集	测试集
FB15k	1345	14904	472860	48991	57803
WN18	18	40943	141442	5000	5000

对于 FB15k 的描述数据集，每段描述的平均词长为 69，最长描述包含 343 个词。对于 WN18 的描述数据集，每段描述的平均词长为 13，最长描述包含 96 个词。

4.2 实验设置

4.2.1 对比模型

对比模型分为三类：（1）本文提出的模型：J-CNN、JA-CNN 和 JPA-CNN；（2）仅利用结构信息的知识表示模型：TransE[1]、Unstructured[7]、SME(linear)[9]、SME(Bilinear)[9]、TransH[10]、TransR[11]、CTransR[11]、TransD[12]和 TranSparse[13]；（3）引入文本信息的表示模型：CNN+TransE[4]、Jointly(LSTM)[5]和 Jointly(A-LSTM)[5]。

4.2.2 参数设置

参数设置：最大间隔 $\gamma \in \{0.1,1,2,5,10\}$，向量维度 $d \in \{50, \}$，学习率 $\lambda \in \{0.000001, 0.0\}$，卷积层的窗口大小 $k \in \{1,2,3,4,5\}$，滤波器的数量 $nf \in \{16,64,128\}$，Dropout 层的丢弃率统一设置为 0.5，不相似性度量 L 设置为 L_1 或 L_2。为加速收敛，本文使用 TransE 的结果来初始化实体和关系的向量。

实验中 J-CNN、JA-CNN 和 JPA-CNN 共享同一组最优参数。在链路预测任务中，模型的最优参数为：$\gamma = 2$，$d = 100$，$\lambda = 0.0001$，$k = 4$，$nf = 64$，$L = L_1$。针对 WN18 数据，模型的最优参数为：$\gamma = 5$，$d = 50$，$\lambda = 0.0001$，$k = 4$，$nf = 64$，$L = L_1$。在三元组分类任务中，针

对 FB15k 数据，模型的最优参数为：$\gamma=1$， $d=100$， $\lambda=0.1$， $k=1$， $nf=16$， $L=L_1$。针对 WN18 数据，模型的最优参数为：$\gamma=0.1$， $d=50$， $\lambda=0.0001$， $k=4$， $nf=64$， $L=L_1$。

本文与 Xu 等人提出的模型[5]在同样的任务上使用相同的数据集和对比模型，因此本文直接使用该论文里对比模型的最优结果进行比较。

4.3 链路预测

链路预测任务旨在预测三元组中缺失的头实体或尾实体。对每个合法三元组，本文会先破坏它的头或者尾实体，依次替换成实体集中的其它实体，然后计算被破坏的三元组的得分，对得分进行升序排序，最后记录三元组的排名。类似 TransE，本文采用两种评估指标：（1）mean rank：所有合法三元组里实体排名的平均值。（2）hits@10：所有合法三元组里实体排名小于 10 的比例。好的表示模型在该任务下应拥有较低的 mean rank 和较高的 hits@10。

评估设置分为两种："原始"（Raw）和"过滤"（Filt）。三元组替换了头或者尾实体后也可能合法，排序时这类破坏的三元组可能会排在合法三元组前面，这并不合理，所以应在排序前删除训练集、测试集和验证集里这类错误三元组，该设置称为"过滤"。本文会展示这两种设置的评估结果。在数据集 WN18 和 FB15k 的实验结果如表 2。

表 2 链路预测的结果

数据集	**WN18**				**FB15k**			
评估指标	**Mean Rank**		**Hits@10**		**Mean Rank**		**Hits@10**	
	Raw	**Filt**	**Raw**	**Filt**	**Raw**	**Filt**	**Raw**	**Filt**
Unstructured[7]	315	304	35.3	38.2	1074	979	4.5	6.3
SME (linear)[9]	545	533	65.1	74.1	274	154	30.7	40.8
SME (Bilinear)[9]	526	509	54.7	61.3	284	158	31.3	41.3
TransH[10]	318	303	75.4	86.7	212	87	45.7	64.4
TransR[11]	238	225	**79.8**	92.0	198	77	48.2	68.7
TransD[12]	224	212	79.6	**92.2**	194	91	**53.4**	**77.3**
CNN+TransE[4]	-	-	-	-	181	91	49.6	67.4
Jointly(LSTM)[5]	117	95	79.5	91.6	179	90	49.3	69.7
Jointly(A-LSTM)[5]	134	123	78.6	90.9	167	73	52.9	75.5
TransE[1]	263	251	75.4	89.2	243	125	34.9	47.1
J-CNN	110	100	77.8	89.3	180	69	52.2	72.3
JA-CNN	**105**	**94**	78.8	90.1	**166**	**68**	53.0	74.7
JPA-CNN	**105**	**94**	77.6	88.9	170	**68**	52.1	71.7

在所有数据集上，本文的模型 JA-CNN 在各项指标都与目前最好的模型 Jointly(A-LSTM) 水平相近，并在 mean rank 指标上达到目前最好的效果，这表明 JA-CNN 能有效捕捉文本的语义信息，在融合多源信息方面有一定优势。

与同样基于 CNN 编码文本的模型 CNN+TransE 相比，本文提出的三个模型在所有指标都有明显提高，可能原因在于：本文选取的 CNN 结构更适应于编码描述的语义信息；本文引入了门机制，加强了两种信息源间的语义联系，比单纯的加权效果要更好；本文设计的注意力机制筛选文本的有效信息，增强了实体表示的区分度。

与仅利用结构信息里最好的模型 TransD 相比，本文的模型在 mean rank 指标取得了大幅度的提升，这表明引入文本信息确实有效缓解数据稀疏，但可能会影响训练过程中的频繁实体，导致 hits@10 指标变差。虽然 hits@10 比 TransD 的表现更差，但考虑到本文是基于 TransE 的改进而不是 TransD。若基于其它优秀的表示模型如 TransD 等进行拓展，应该能进一步提升模型性能。

本文的三种模型互相对比，JA-CNN 的性能优于 J-CNN，这表明注意力机制的引入加强文本表示的语义区别，进一步提高了实体表示的区分度。在 FB15k 数据集上，JPA-CNN 比 JA-CNN 表现要逊色许多，这可能是因为在该数据集上句子的长度长短分化严重，固定的位置编码没法有效拟合出这种差异，反而一定程度上成为干扰信息。在 WN18 数据集上，JPA-CNN 与 JA-CNN 效果相近，比在 FB15k 上表现更好，这可能是因为该数据集的句子整体偏短，长度差异较小，本文的位置编码更适用于拟合该数据集。

4.3.1 关系分类实验

为进一步展示模型的性能，本文把关系划分成 1-1、1-N、N-1 和 N-N 四种类型，并比较不同类型关系下模型在数据集 FB15k 上 hits@10（Filt）的结果。

表 3 的结果表明，面对所有类型的关系，本文的模型 JA-CNN 都比对比模型展现了更好的性能，尤其在 N-1 和 N-N 关系下的头实体预测以及 1-N 和 N-N 的尾实体预测上有明显的提高，这表明 JA-CNN 面对知识库的复杂关系时有一定优势。

表 3 关系分类的结果

任务	**头实体预测 (Hits@10)**				**尾实体预测 (Hits@10)**			
关系类型	**1-1**	**1-N**	**N-1**	**N-N**	**1-1**	**1-N**	**N-1**	**N-N**
Jointly(A-LSTM)[5]	83.8	95.1	21.1	47.9	83.0	30.8	**94.7**	53.1
TransE[1]	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
J-CNN	82.5	94.7	35.9	72.1	80.2	45.4	94.6	75.3
JA-CNN	**84.8**	**95.5**	**39.7**	**74.5**	**84.8**	**49.7**	**94.7**	**77.8**
JPA-CNN	81.4	95.2	36.8	71.2	82.8	45.4	94.5	74.5

4.4 三元组分类

三元组分类旨在判断给定的三元组是否合法。本文使用 FB15k 和 WN18 的数据集来评估模型的性能。由于数据集只有正样本，本文按照 Socher 等人的方法[17]来构造负样本。该方法随机替换合法三元组的头实体来构成负样本，替换的实体只能从该三元组的关系对应的实体集中选择。该方法使负例集合里不会出现明显的无关系三元组，使负例与正例的语义差别更小，进而增加评估任务的难度。

本文使用准确率作为该任务的评估指标。任务先达到验证集的最大准确率，获得每个关系 r 的阈值 δ_r；接着对测试集的每个三元组 (h,r,t) 计算得分，若三元组的得分小于 δ_r，会归为合法三元组，否则归为错误三元组。在 FB15k 数据集上，有部分关系出现在验证集却没有出现在测试集中，本文会采用验证集中出现过的关系的阈值的平均值来补充缺失的阈值。表 4 展示了三元组分类的结果。

表 4 三元组分类的结果

数据集	WN18	FB15k
TransH[10]	-	79.9
TransR[11]	-	82.1
CTransR[11]	-	84.3
TransD[12]	-	88.0
TranSparse[13]	-	88.5
Jointly(A-LSTM)[5]	97.8	**91.5**
TransE[1]	92.9	79.8
J-CNN	96.8	87.3
JA-CNN	97.8	87.0
JPA-CNN	**98.0**	90.1

结果表明本文的联合模型相比 TransE 有大幅度提高，其中，JPA-CNN 与最好的模型性能相近，这表明本文的文本编码方法能有效编码语义信息，并能很好的融合到实体表示，进而增强了三元组的语义区分度。

本文的三种模型相比，JPA-CNN 均达到了最佳性能，这表明引入位置信息后，面对复杂关系能找到更精准的阈值，同时增强正负样本得分的差距。在 FB15k 上，JA-CNN 比 J-CNN 效果稍差，但在 WN18 中，JA-CNN 比 J-CNN 效果更好，这可能是因为 FB15k 数据集里描述的句子整体偏长，且关系的种类远大于 WN18，本文的注意力机制在模拟这种语义信息时，一定程度上弱化了三元组得分的差异。

5 总结

本文提出了基于带注意力机制 CNN 的联合知识表示模型，并通过引入实体描述信息来改善知识表示。首先，本文提出了基于 CNN 的文本编码器；然后，设计相应的注意力机制来筛选与关系最相关的文本信息；接着，又引入位置信息拓展该模型；最后，利用门机制联合文本信息和结构信息获得最终的联合表示。实验证明，本文的模型在链路预测和三元组分类任务上与目前最好的模型水平相近，并在关系分类任务上表现更好，这表明本文的方法能有效融合多源数据，缓解知识库稀疏，也为改善实体表示提供了新思路。

未来，本文将考虑以下方向来改进模型：

（1）本文采用了基于 TransE 的得分函数，未来可以考虑基于其它优秀的知识表示模型如 TransD 等进行模型拓展。

（2）考虑融合更多其它信息例如实体的类别信息。本文会尝试拓展模型来编码类别信息，也考虑使用类别作为实体的约束信息。

参考文献

[1]Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]. In NIPS’13, 2013: 2787-2795.

[2]Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph and text jointly embedding[C]. In EMNLP’14, 2014: 1591–1601.

[3]Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. Aligning knowledge and text embeddings by entity descriptions[C]. In EMNLP’15, 2015: 267–272.

[4]Ruobing Xie, Zhiyuan Liu, Jia Jia, et al. Representation Learning of Knowledge Graphs with Entity Descriptions[C]. In AAAI’16, 2016: 2659-2665.

[5]Jiacheng Xu, Xipeng Qiu, Kan Chen, et al. Knowledge graph representation with jointly structural and textual encoding[C]. In IJCAI’17, 2017: 1318-1324.

[6]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. In NIPS’17, 2017: 6000-6010.

[7]Antoine Bordes, Xavier Glorot, Jason Weston, et al. Joint learning of words and meaning representations for open-text semantic parsing[C]. In AISTATS’12, 2012: 127-135.

[8]Antoine Bordes, Jason Weston, Ronan Collobert, et al. Learning Structured Embeddings of Knowledge Bases[C]. In AAAI’11, 2011: 301–306.

[9]Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(2): 233-259.

[10]Zhen Wang, Jianwen Zhang, Jianlin Feng, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]. In AAAI’14, 2014: 1112-1119.

[11]Yankai Lin, Zhiyuan Liu, Maosong Sun, et al. Learning entity and relation embeddings for knowledge graph completion[C]. In AAAI’15, 2015: 2181-2187.

[12]Guoliang Ji, Shizhu He, Liheng Xu, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[C]. In ACL'15, 2015: 687–696.

[13]Guoliang Ji, Kang Liu, Shizhu He, et al. Knowledge Graph Completion with Adaptive Sparse Transfer Matrix[C]. In AAAI'16, 2016: 985-991.

[14]Han Xiao, Minlie Huang, Yu Hao, et al. TransA: An adaptive approach for knowledge graph embedding[J]. arXiv preprint arXiv:1509.05490, 2015.

[15]Shizhu He, Kang Liu, Guoliang Ji, et al. Learning to repre-sent knowledge graphs with gaussian embedding[C]. In CIKM'15, 2015: 623-632.

[16]Han Xiao, Minlie Huang, Xiaoyan Zhu. TransG: A genera-tive model for knowledge graph embedding[C]. In ACL'16, 2016: 2316-2325.

[17]Richard Socher, Danqi Chen, Christopher D Manning, et al. Reasoning with neural tensor networks for knowledge base completion[C]. In NIPS'13, 2013: 926-934.

[18]Dongxu Zhang, Bin Yuan, Dong Wang, et al. Joint semantic relevance learning with text data and graph knowledge[C]. In ACL-IJCNLP'15, 2015: 32-40.

[19]Ni Lao, Amarnag Subramanya, Fernando Pereira, et al. Reading the web with learned syntactic-semantic inference rules[C]. In EMNLP-CoNLL'12, 2012: 1017-1026.

[20]Toutanova K, Chen D, Pantel P, et al. Representing text for joint embedding of text and knowledge bases[C]. In EMNLP'15, 2015: 1499-1509.

[21]Arvind Neelakantan, Benjamin Roth, Andrew McCallum. Compositional Vector Space Models for Knowledge Base Completion[C]. In ACL'15, 2015: 156–166.

[22]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[23]Sukhbaatar S, Weston J, Fergus R. End-to-end memory net-works[C]. In NIPS'15, 2015: 2440-2448.

[24]Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[25]Bollacker K, Evans C, Paritosh P, et al. Freebase: a collabora-tively created graph database for structuring human knowledge[C]. In SIGMOD'08, 2008: 1247-1250.

作者联系方式：彭敏　湖北省武汉市武昌区八一路 299 号武汉大学计算机学院 430072
18602719688 pengm@whu.edu.cn