基于改进注意力机制的实体关系抽取方法

冯建周[1],宋沙沙[1],王元卓[1,2],刘亚坤[1],武红颖[1],龚 昊[1]
(1. 燕山大学信息科学与工程学院,河北秦皇岛 066000; 2. 中国科学院计算技术研究所,北京 100080)

摘 要: 实体关系抽取是知识库构建中至关重要的一个环节. 在众多的实体关系抽取方法中,远程监督结合神经网络模型的方法在准确率等性能上是比较令人满意的,但远程监督获取的标注语料中往往存在大量的噪声数据,给实体关系抽取模型的训练带来了很大的影响. 本文提出一种基于改进注意力机制的卷积神经网络实体关系抽取模型. 该模型针对包含同一实体对的句子集合,从中尽可能地找出所有体现该实体对关系的正实例,构建组合句子向量,抛弃可能的噪声句子,从而最大程度地降低噪声句子的影响又能充分利用正实例的语义信息. 实验证明,本文提出的关系抽取模型在准确率上优于对比的关系抽取模型.

关键词: 关系抽取; 改进注意力机制; 卷积神经网络; 远程监督; 组合句子特征向量
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2019)08-1692-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.08.012

Entity Relation Extraction Based on Improved Attention Mechanism

FENG Jian-zhou[1], SONG Sha-sha[1], WANG Yuan-zhuo[1,2], LIU Ya-kun[1], WU Hong-ying[1], GONG Hao[1]
(1. *School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei* 066000, *China*;
2. *Institute of Computer Technology, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: Entity relation extraction is a crucial part of knowledge base construction. Among many methods of relationship extraction, the method of distant supervision combined with neural network model is satisfactory in terms of accuracy and other performance. However, there is often a large amount of noise data in the labeled corpus obtained by distant supervision, which has a great impact on the training of relationship extraction model. In this paper, we propose an entity relationship extraction model of convolutional neural network based on improved attention mechanism. Aiming at the sentence set containing the same entity pair, this model tries to find out all the positive instances that embody the relationship between the entity pair, construct the combined sentence vector, and discard the possible noise sentences, so as to minimize the impact of noise sentences and make full use of the semantic information of positive instances. Experimental results show that the accuracy of the proposed relation extraction model is better than that of the comparative relation extraction model.

Key words: relation extraction; improved attention mechanism; convolutional neural networks; distant supervision; combined sentence feature vector

1 引言

关系抽取是信息抽取(Information Extraction, IE)[1]的主要任务之一,又称实体关系抽取,是指对文本中所含有实体对进行语义关系分类,在智能问答、知识库构建等领域扮演着重要的角色. 关系抽取的研究方法众多,其中,基于机器学习的关系抽取方法应用最为广泛. 机器学习的方法根据是否需要标注好的训练语料可分为有监督的学习方法、无监督的学习方法[2~4]和半监督的学习方法[5~9]. 其中有监督的学习方法[10~12]性能最好,但是需要大量的人工标注语料,耗时费力,于是远程监督关系抽取方法应运而生. Mintz 等人[13]尝试使用 FreeBase 知识库来代替手工标注语料,他们利用 Freebase 与自由文本对齐得到大量的标注训练语料,Mintz 称这种监督方法为远程监督.

远程监督的思想基于一种假设:如果两个实体之间存在知识库中的某种关系,那么含有这两个实体的句子或多或少都表达了这种关系. 基于这种假设,远程

收稿日期:2017-11-27;修回日期:2019-03-22;责任编辑:蓝红杰
基金项目:国家自然科学基金青年基金(No. 61602401);河北省教育厅高等学校科技计划青年基金项目(自然类)(No. QN2018074)

监督的关系抽取方法把包含同一实体对的句子都作为该实体对在知识库中所对应关系的正例. 然而远程监督的关系抽取方法仍存在两个关键问题:首先,该方法仍然需要利用自然语言处理(Natural Language Processing,NLP)工具获取手工设计句子特征,不可避免地存在一些错误,影响关系抽取的效果;其次,该方法获得的实体关系标注语料中仍然存在大量的噪声数据. 为了解决这两个问题,文献[14]提出了一种基于多实例学习的分段卷积神经网络(Piecewise Convolutional Neural Networks,PCNN)模型. 该模型既能自动提取句子特征,又能抛弃大量的噪声句子. 但是,该模型对于每个实体对只选用一条句子进行学习和预测,损失了来自其他句子的大量的有效信息. 文献[15~17]引入注意力机制对每条句子赋予权重,其中正实例赋予较高权重,噪声句子赋予较低权重,使得能够在利用更多句子语义信息的同时,尽可能降低噪声句子的影响. 尽管在注意力机制中,噪声句子获得的权重比正实例句子的权重小,但仍会对模型的性能有影响,且包含同一实体对的句子集中噪声句子越多,这种影响越大. 假设包含同一实体对的句子集中有 10 条句子,其中三条句子是正实例,剩下的七条句子为噪声句子,尽管每条噪声句子的权重都比较小,但七条噪声句子的权重加起来对模型性能的损害也是相当严重的.

本文提出一种基于改进注意力机制(Imporved sentence-level ATTention,IATT)的卷积神经网络实体关系抽取模型. 该模型的贡献主要体现在以下几点:(1)提出了一种新的注意力模型,区别于之前建立在单个句子向量上的注意力机制,该模型以组合句子向量为基本单位设计注意力机制,从而可以更加彻底的抛弃噪声句子,而不是让噪声句子以一个较低权重参与计算;(2)以组合句子向量为基本单位的注意力机制,能够最大限度的保留正实例句子,从而可以丰富该实体对关系的语义信息,提高模型的泛化能力;(3)设计对比实验,验证了本文所提的新的注意力模型在性能上的优势.

2　基于改进注意力机制的卷积神经网络模型

本文提出一种基于改进注意力机制的卷积神经网络关系抽取模型(IATT),在句子级注意力机制的基础上,不再采用降低噪声句子权重的方法,而是彻底剔除噪声句子,通过构建排序的组合向量的方法,找到不包含噪声句子的组合向量,从而使得模型的训练效果更好. 如图 1 所示.

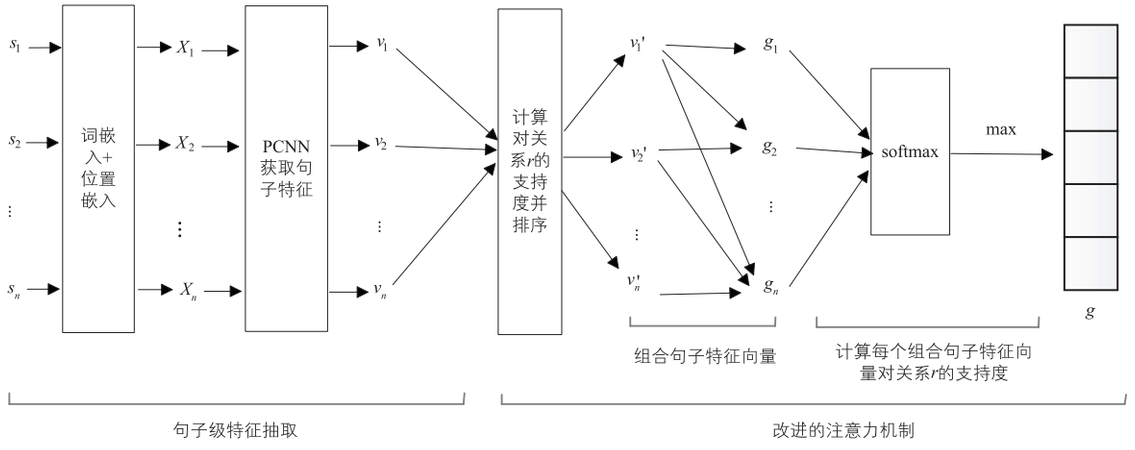

图1　基于改进注意力机制的卷积神经网络模型

对于包含相同实体对的句子集合,该模型首先利用词向量嵌入技术和卷积神经网络得到每条句子的特征向量表示$\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$,然后基于每条句子的特征向量计算其对标签关系的支持度,并基于计算出的支持度对句子的特征向量集合进行排序,得到有序的句子特征向量集合$\{\boldsymbol{v}'_1,\boldsymbol{v}'_2,\cdots,\boldsymbol{v}'_n\}$,然后从得分最高的句子特征向量开始递增式地组合多条句子特征向量,形成组合特征向量集合$\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_n\}$,$\boldsymbol{g}_1=\boldsymbol{v}'_1$,$\boldsymbol{g}_2=\boldsymbol{v}'_1\boldsymbol{v}'_2$,$\cdots$,$\boldsymbol{g}_n=\boldsymbol{v}'_1\boldsymbol{v}'_2\boldsymbol{v}'_3,\cdots,\boldsymbol{v}'_n$,其中$\boldsymbol{v}'_1$是得分最高的句子特征向量,$\boldsymbol{v}'_1\boldsymbol{v}'_2$表示得分最高和得分次高的两个句子特征向量组成的组合特征向量,以此类推,$\boldsymbol{v}'_1\boldsymbol{v}'_2\boldsymbol{v}'_3,\cdots,\boldsymbol{v}'_n$是包含同一实体对的所有句子特征向量组合而成的组合特征向量. 最后通过计算每个组合特征向量在多分类输出时映射到标签关系 r 上的得分,得到得分最高的组合特征向量,即认为该组合特征向量是包含了最多正实例语义信息、最少噪声句子信息的组合特征向量,并利用该组合特征向量为目标实体对训练分类器.

图 1 中,$\{s_1,s_2,\cdots,s_n\}$代表包含同一实体对的句子集合,$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$分别代表每条句子的向量表示,$\boldsymbol{g}$为在标签关系 r 上得分最高的组合句子特征向量,即包

含了最多的正实例语义信息最少的噪声句子信息的组合句子特征向量.

本文将从两部分描述基于改进注意力机制的卷积神经网络模型:句子级特征抽取,改进的注意力机制.

2.1 句子级特征抽取

句子级特征抽取的目的是将句子信息转换成特征向量. 如图 2 所示,首先要把句子转换成向量表示,然后通过卷积神经网络的卷积操作和池化操作得到句子向量的特征向量.

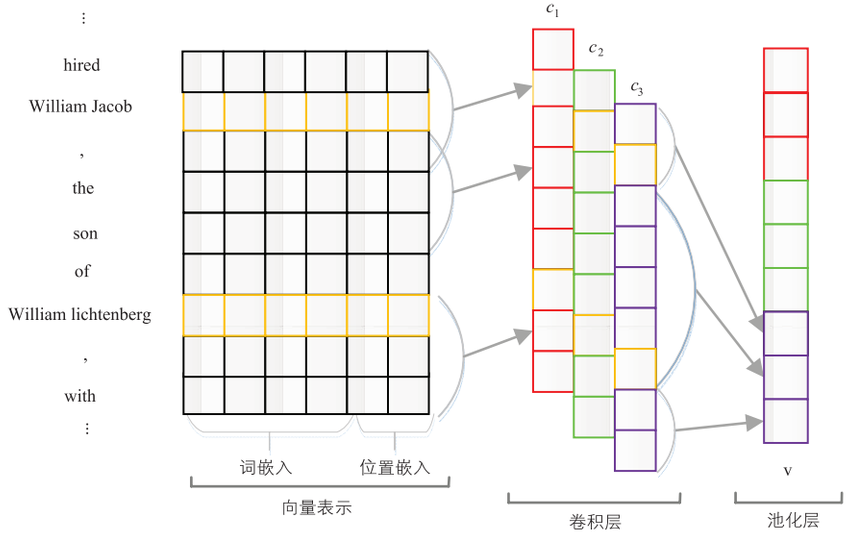

图2 句子级特征抽取结构图

2.1.1 向量表示

自然语言文本的句子不能被神经网络直接编码,所以利用神经网络处理 NLP 任务时要把自然语言文本中的句子转化为向量表示. 句子的向量表示包括词嵌入和位置嵌入两部分.

(1)词嵌入

本文采用将句子中的每个单词转化成 K 维实值向量构建模型的输入. 给定一个句子 $\boldsymbol{s}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_m)$,通过映射词向量矩阵 $\boldsymbol{E}\in\mathbf{R}^{|V|\times d_w}$ 将每个单词 $\boldsymbol{w}_i$ 表示为 d_w 维实值向量,V 是词表的大小(词向量训练语料中的词的数目).

(2)位置嵌入

词嵌入虽然能很好地捕捉到单词的词义信息,但它无法捕捉句子的结构信息. 在判断句子中两个目标实体的关系时,距离实体近的单词通常是关键的信息,所以指出句子中目标实体的位置是必要的. 本文利用位置特征来记录当前单词到两个实体的相对距离[14],例如:在句子“Steve Jobs was born in San Francisco.”中,单词 born 到头实体 Steve Jobs 的相对距离是 2,到尾实体 San Francisco 的相对距离是 -2. 然后将这两个相对距离映射成随机初始化的两个 d_p 维的实值向量.

对于每条句子将其每个单词的词嵌入和位置嵌入连接起来就得到该句子的向量表示矩阵 $\mathbf{X}\in\mathbf{R}^{m\times d}$,其中,$m$ 表示句子的长度,d 是单词嵌入和位置嵌入连接后的维度,即 $d=d_w+d_p\times 2$.

2.1.2 卷积操作

在实体关系抽取任务中,每个句子的长度是不统一的,并且判断目标实体间关系类别的重要信息可能分布在句子的任何地方,这就意味着必须要从句子中抽取不同的局部特征来预测目标实体对所属的关系类型. 在神经网络中,卷积操作是获取这些局部特征的常用方法[18]. 卷积操作被定义为权重矩阵 $\boldsymbol{W}\in\mathbf{R}^{w\times d}$(又叫过滤器)和句子向量表示 $\boldsymbol{X}$ 之间的操作. 卷积操作首先使用大小为 w 的滑动窗口(窗口大小即过滤器的大小)与每个句子的向量表示进行卷积计算. 窗口滑动到句子结尾处超出索引边界时用 $\mathbf{0}$ 向量来填充. 把句子向量表示 $\boldsymbol{X}$ 看作一个序列$(\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_m)$,其中 $\boldsymbol{q}_i\in\mathbf{R}^d$,$\boldsymbol{q}_{i:j}$ 表示 $\boldsymbol{q}_i$ 到 $\boldsymbol{q}_j$ 的连接,卷积计算其实就是权重矩阵 $\boldsymbol{W}$ 与 w 个序列 $\boldsymbol{q}_{i:i+w-1}$ 进行点积运算得到一个序列 $\boldsymbol{c}\in\mathbf{R}^{m+w-1}$:

$$\boldsymbol{c}_j=\boldsymbol{W}\boldsymbol{q}_{j-w+1:j}+\boldsymbol{b},\quad 1\leqslant j\leqslant m+w-1 \tag{1}$$

其中,$\boldsymbol{b}$ 是偏置向量.

由于要抽取句子向量的多个局部特征,所以需要用多个不同的过滤器来完成局部特征抽取,假设设置了 n 个过滤器 $\boldsymbol{W}=(\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_n)$,则每个过滤器都得到一个特征向量:

$$\boldsymbol{c}_{ij}=\boldsymbol{W}_i\boldsymbol{q}_{j-w+1:j},\quad 1\leqslant i\leqslant n \tag{2}$$

最终卷积层输出一个矩阵 $\boldsymbol{C}=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n)\in\mathbf{R}^{n\times(m+w-1)}$,其中,$n$ 为过滤器个数,m 为句子的长度,w

为窗口大小.

2.1.3 分段池化操作

卷积层输出向量 $\boldsymbol{C}$ 的大小依赖于输入神经网络的句子向量矩阵 $\boldsymbol{X}$ 的长度,为了使特征向量独立于句子的长度,需要池化层来进行特征压缩和进一步提取主要特征. 卷积神经网络中通常采用最大池操作进一步提取特征,该操作对于每个过滤器抽取到的若干特征值,只取其中最大的那个值作为池化层的保留值,其它特征值全部抛弃,即同一类特征中只保留最强的,而抛弃其它弱的此类特征. 为了更加细腻地捕捉句子中两个实体之间的结构信息,本文采用分段最大池化进一步提取特征[14].

对于从卷积层得到的 n 个特征向量($\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n$),将每个特征向量 $\boldsymbol{c}_i$ 以头实体和尾实体为界分为三段($\boldsymbol{c}_{i1},\boldsymbol{c}_{i2},\boldsymbol{c}_{i3}$),分别从每段 $\boldsymbol{c}_{ij}$ 中捕捉最大的值.

$$p_{ij}=\max\{\boldsymbol{c}_{ij}\},\ 1\leqslant i\leqslant n,1\leqslant j\leqslant 3 \tag{3}$$

每一个过滤器都得到一个三维向量 $\boldsymbol{p}_i=(p_{i1},p_{i2},p_{i3})$,将所有的特征向量连接起来 $\boldsymbol{p}_{1:n}$,最后利用非线性函数,例如双曲正切函数,生成一个向量 $\boldsymbol{v}$:

$$\boldsymbol{v}=\tanh(\boldsymbol{p}_{1:n}),\ \mathbf{v}\in\mathbf{R}^{3n} \tag{4}$$

最终向量 $\boldsymbol{v}$ 为句子 s 的特征向量表示.

2.2 改进的句子级注意力机制

为了过滤掉远程监督中的大量噪声句子,本文改进了句子级注意力机制,尽可能抛弃所有负实例句子,将所有正实例句子的特征组合起来作为训练正例.

2.2.1 构造基于权重有序的句子向量集合

假设包含实体对 $<e_1,e_2>$(标注关系为 r)的句子集合 S 中有 n 条句子 $S=\{s_1,s_2,\cdots,s_n\}$,经由 2.1 节可以得到每条句子的特征向量表示,构成向量集合 $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$. 句子集合中每条句子表达关系 r 的程度是不一样的,因此为每条句子设置一个权重反映该句子对关系 r 的支持度. 计算每条句子的权重($\beta_1,\beta_2,\cdots,\beta_n$)的公式如下:

$$\beta_i=\frac{\exp(e_i)}{\sum_{k=1}^{n}\exp(e_k)},\quad 1\leqslant i\leqslant n \tag{5}$$

其中,e_i 是句子集合中第 i 条句子与标签关系 r 的相关度,其计算方法如下:

$$e_i=\boldsymbol{v}_i\boldsymbol{A}\boldsymbol{z},\quad 1\leqslant i\leqslant n \tag{6}$$

其中,$\boldsymbol{v}_i$ 代表句子集合中第 i 条句子的特征向量,$\boldsymbol{A}$ 是一个对角权重矩阵,$\boldsymbol{z}$ 是关系 r 的查询向量,代表关系 r. 计算完每条句子的权重后,将句子特征向量集合 $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$ 按照权重高低排序为 $\{\boldsymbol{v}'_1,\boldsymbol{v}'_2,\cdots,\boldsymbol{v}'_n\}$.

2.2.2 生成组合句子特征向量

本文认为由越多的正实例组合的句子特征在标签关系上的得分越高,而掺杂越多的负实例组合的句子特征在标签关系上的得分越低. 基于以上假设,本文试图通过学习找到在标签关系上的得分最高的组合句子特征向量. 首先,用基于权重有序的句子特征向量集合 $\{\boldsymbol{v}'_1,\boldsymbol{v}'_2,\cdots,\boldsymbol{v}'_n\}$ 生成组合句子特征向量集合 $\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_n\}$. $\boldsymbol{g}_i$ 的生成方法如下:

$$\boldsymbol{g}_i=\sum_{j=1}^{i}\alpha_j\boldsymbol{v}'_j,\ 1\leqslant i\leqslant n \tag{7}$$

其中 α_j 的计算公式如下:

$$\alpha_j=\frac{\exp(e_j)}{\sum_{k=1}^{i}\exp(e_k)},\ 1\leqslant j\leqslant i \tag{8}$$

其中,e_j 是句子集合中第 j 条句子与标签关系 r 的相关度,其计算方法按照式(6).

2.2.3 softmax 分类器

以上构造的 n 个组合句子特征向量 $\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_n\}$ 中,必然有一个组合特征向量包含了尽可能多的正实例句子信息,尽可能少的噪声句子信息,这样的特征向量作为实体对 $<e_1,e_2>$ 的训练正例是最佳的. 本文通过计算每个组合特征向量在关系 r 上的条件概率来选择最佳的组合句子特征向量.

首先计算每个组合句子特征向量在关系 r 上的得分,将上述得到的 n 个组合句子特征向量 $\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_n\}$ 依次输入到 softmax 分类器.

$$\boldsymbol{o}_i=\boldsymbol{W}_0\boldsymbol{g}_i+\boldsymbol{b},\quad 1\leqslant i\leqslant n \tag{9}$$

其中,$\boldsymbol{W}_0\in\boldsymbol{R}^{h\times 3n}$ 是一个权重矩阵,h 是关系抽取系统预定义关系类型的数目. $\boldsymbol{o}_i$ 是组合句子特征向量 $\boldsymbol{g}_i$ 在 softmax 上的输出向量.

为了防止过拟合,本文在 softmax 分类器中加入 dropout 算法[19],dropout 是指在训练时,按一定的概率 p 来对权重层的参数进行随机采样,将这个子网络作为此次更新的目标网络.

在训练时给向量 $\boldsymbol{g}_i$ 添加操作($\boldsymbol{g}_i\circ\boldsymbol{f}$),式(9)变成如下公式:

$$\boldsymbol{o}_i=\boldsymbol{W}_0(\boldsymbol{g}_i\circ\boldsymbol{f})+\boldsymbol{b},\ 1\leqslant i\leqslant n \tag{10}$$

其中,$\boldsymbol{W}_0\in\mathbf{R}^{h\times 3n}$ 是一个权重矩阵,h 是关系抽取系统预定义关系类型的数目,$\boldsymbol{f}$ 是概率为 p 的伯努利分布产生的向量. $\boldsymbol{o}_i$ 是组合句子特征向量 $\boldsymbol{g}_i$ 在 softmax 上的输出向量.

2.2.4 组合特征选择

组合特征向量 $\boldsymbol{g}_i$ 经过卷积神经网络后的输出 $\boldsymbol{o}_i$,用 softmax 函数归一化后,可以得到一个条件分布函数 $p(r|\boldsymbol{g}_i;\theta)$,它表示在已知参数 θ 时将特征向量 $\boldsymbol{g}_i$ 归为关系 r 的概率.

$$p(r\mid\boldsymbol{g}_i;\theta)=\frac{\exp(\boldsymbol{o}_i^r)}{\sum_{k=1}^{h}\exp(\boldsymbol{o}_i^k)},\quad 1\leqslant i\leqslant n \tag{11}$$

其中,h 是实体关系抽取系统预定义关系类型的数目,o_i^r 是组合句子特征向量 g_i 在关系 r 上的得分,θ 代表模型的参数(详见 2.3 节).

最终,最佳组合特征向量 g 按式(12)选择:

$$g = \arg\max p(r|g_i;\theta),\ 1 \leqslant i \leqslant n \tag{12}$$

其中,r 代表标签关系,θ 代表模型的参数. 图 3 为改进的注意力机制结构图.

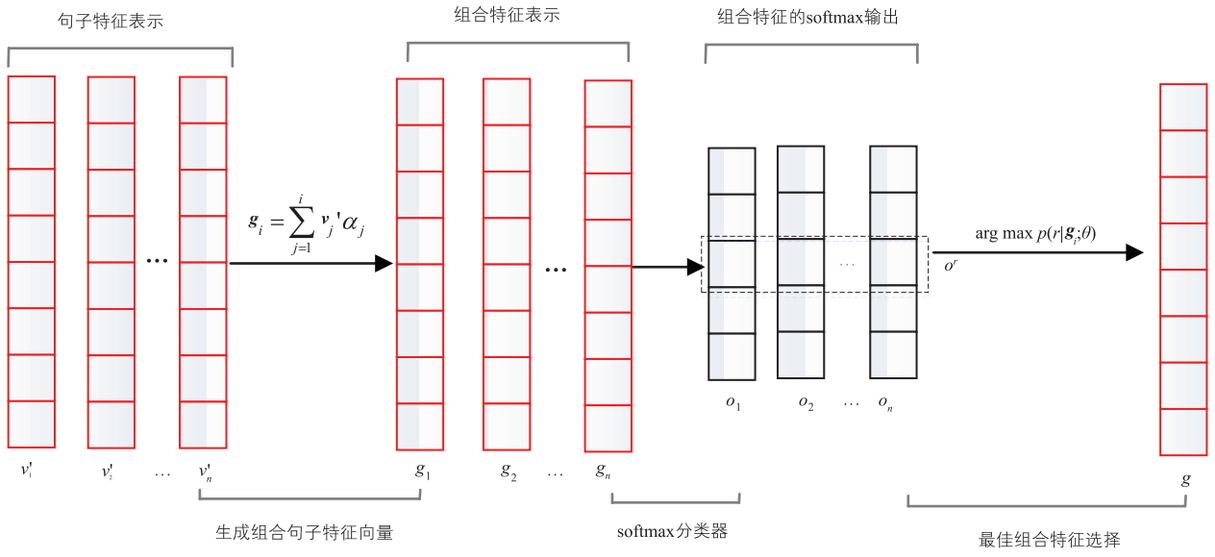

图3 改进的注意力机制结构图

图 3 中,$\{v_1', v_2', \cdots, v_n'\}$ 代表基于权重有序的句子特征向量集合,$\{g_1, g_2, \cdots, g_n\}$ 代表组合后的句子特征集合,$\{o_1, o_2, \cdots, o_n\}$ 代表对应组合特征向量的 softmax 输出,o_i^r 代表 o_i 在关系 r 上分量,即对应组合特征向量在关系 r 上的得分.

2.3 优化目标函数

本文模型需要优化的参数 $\theta = (E, T_{PF1}, T_{PF2}, W, W_0)$,其中,$E$ 代表词向量,T_{PF1}、T_{PF2} 分别为代表单词相对头实体和尾实体的距离的位置向量,W 代表卷积操作的参数,W_0 代表分类器的参数. 在句子集合 S 上利用交叉熵定义目标函数如下:

$$J(\theta) = \sum_{j=1}^{N} \log p(r_j \mid S_j^i, \theta) \tag{13}$$

其中,N 代表句子集合的数量,S_j^i 代表第 j 个句子集合中的最佳组合句子特征向量,根据式(12)计算获得.

3 实验与评估

为了证明本文提出算法的优越性,本节设置了几组对比实验,从不同的角度论证本文算法的优势.

3.1 数据集

本文使用的数据集是文献[14]过滤版本的 NYT10 数据集. 原始的 NYT10 数据集由 Riedel 等人[20]发布并被许多远程监督关系抽取研究[21,22]使用. 该数据集是纽约时报语料库对齐 Freebase 中的关系产生的,从 2005 ~ 2006 年的新闻语料中获取的句子作为训练集,从 2007 年的新闻语料中获取的句子作为测试集. 文献[14]在 Riedel 版本的 NYT10 数据集的基础上过滤掉了以下几种类型的句子:(1)对每个实体对而言重复的句子;(2)两个实体之间的单词数超过 40 的句子;(3)句子中实体名称是 Freebase 中其他实体名称子字符串的句子. 通过过滤以上三种句子,一些低频的关系被移除了.

3.2 实验设置

3.2.1 预训练的词向量

本文所用词向量是利用 Google 的开源工具包 word2vec 工具训练纽约时报语料库得到的. word2vec 首先构造训练文本数据的词表(本文把训练文本中出现频率超过 100 次的单词列入词表),然后学习这些词的向量表示. 对于实体是多个词的情况,把单词连接起来看做一个词. 本文实验中使用 50 维的实值向量表示单词的词向量.

3.2.2 参数设置

本文在训练时使用三折交叉验证法[22]调整模型. 用网格搜索法来确定最优参数,并指定参数空间子集为:窗口大小 $w \in \{1,2,3,\cdots,7\}$,过滤器数量 $n \in \{50, 60, \cdots, 300\}$,批大小 $B \in \{40, 160, 640, 1280\}$,随机梯度下降学习率 $\lambda \in \{0.1, 0.01, 0.001, 0.0001\}$,位置向量维度 $d_p = 5$,使用 Adadelta 优化器[23]更新参数. 本文实验使用的参数如表 1 所示.

表 1　实验参数表

窗口大小	特征图	词嵌入维度	位置嵌入维度	批大小	优化器参数	丢弃率	学习率
$w=3$	$n=230$	$d_w=50$	$d_p=5$	$B=160$	$\rho=0.95$, $\varepsilon=1e^{-6}$	$p=0.5$	$\lambda=0.01$

3.3　实验比较

本文采用自动评估(held-out evaluation)[13]方法来评估模型. 自动评估针对人工标注实体关系(NA 关系除外)的测试数据集,将模型测试结果与人工标注进行自动比较,并采用准确率-召回率曲线来衡量,该曲线是将每条句子测试输出的结果按照在每个关系上的输出概率的高低顺序排列,然后逐条分析每条句子在该关系上是否正确,从而得到一条准确率-召回率曲线.

3.3.1　与经典注意力算法的比较

为了证明本文算法的优势,我们将基于 PCNN[14]模型上的 3 种经典注意力算法与本文提出改进注意力模型在 PCNN 上的实现(PCNN + IATT)进行自动评估比较. 这 3 种算法包括:多实例模型(PCNN + MIL)[14],该算法选择标为某种关系的句子集中权重最高的单个句子训练神经网络模型,可以认为是一种最简单的注意力算法;均值注意力模型(PCNN + AVE)[15]和句子级选择性注意力模型(PCNN + ATT)[15]. 图 4 是四种算法的召回率-准确率对比图.

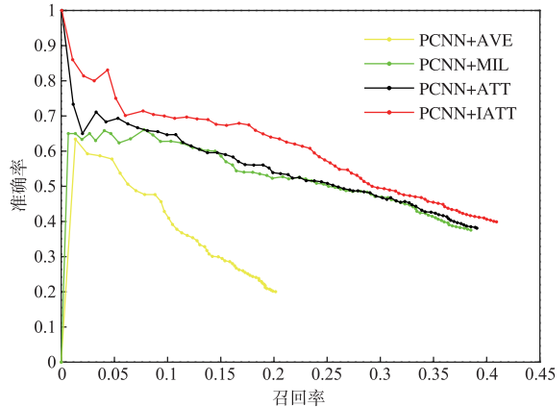

图4　PCNN+AVE, PCNN+MIL, PCNN+ATT, PCNN+IATT的召回率–准确率曲线

从图 4 中可以看出:(1)PCNN + AVE 算法的性能最差甚至不如 PCNN + MIL 算法. 这是因为 PCNN + AVE 算法同等对待句子集合中的每一条句子,虽然该算法比 PCNN + MIL 算法获取到更为丰富的语义信息,但同时也引入了噪声数据的特征;(2)PCNN + ATT 算法的效果要比 PCNN + MIL 算法和 PCNN + AVE 算法好,这是因为它在充分利用更多句子信息的同时,还通过减少噪声句子权重来降低噪声句子对模型训练的影响;(3)PCNN + IATT 算法在整个召回率范围内取得了最高的准确率. 说明 PCNN + IATT 算法的性能是优于所有对比模型的,因为它不但能够充分利用更多的正实例句子信息,还能尽可能的剔除噪声句子,不让噪声句子参与模型训练,而不像 PCNN + ATT 那样只是减小噪声句子参与训练的权重,在去噪声方面做的更彻底.

表 2 是几种算法的 Top$N$ 比较表,该表表示在按照测试输出的概率排序的基础上,前 N 条句子的准确率. 从表 2 中列出的 Top100,Top200 和 Top500 的比较可以看出:(1)PCNN + AVE 算法的准确率都是最低的,这说明噪声对模型的影响严重;(2)PCNN + ATT 算法要全面优于 PCNN + MIL 算法. 说明了选择注意力模型更多的利用句子信息,同时减少了噪声的影响;(3)PCNN + IATT 算法较 PCNN + ATT 算法有 3% 到 11% 的提升,说明本文算法进一步减少了噪声的影响,也进一步验证了本文算法的优势.

表 2　PCNN + IATT 与 PCNN + AVE, PCNN + MIL, PCNN + ATT 的 Top$N$ 对比表

Top$N$	Top100	Top200	Top500
PCNN + AVE	0.600	0.540	0.404
PCNN + ONE	0.630	0.635	0.586
PCNN + ATT	0.700	0.675	0.592
PCNN + IATT	0.810	0.710	0.654

3.3.2　与其他改进的注意力算法的比较

为了进一步体现本文算法的优势,我们选择了基于文献[14,15]改进的 2 种注意力算法进行比较,这 2 种算法包括:句子级注意力加实体描述算法(APCNNs + D)[16]和结合多种注意力机制的 Memory 算法[17]. 图 5 是三种算法比较的准确率-召回率曲线.

图 5 中,APCNN + D 算法是在句子级注意力机制的基础上增加了额外的实体描述信息来进一步降低噪声句子的权重,而 Memory 算法则结合了词注意力、句子级注意力和关系依赖三种注意力机制. 本文所提的 PCNN + IATT 算法只采用一种改进的注意力模型,而且并未利用任何额外信息. 从图 5 可见,在召回率 0 ~ 0.1 范围内,三种算法曲线犬牙交错,各有优劣,在召回率 0.1 以后,本文的算法要明显优于其他两种算法,因此,我们认为本文算法在整体性能上还是取得了优势.

3.3.3　与传统非注意力算法比较

为了充分体现本文算法的优势,我们将本文算法与几种有代表性的传统非注意力算法进行了比较,本文选取 Mintz[13]和 MIML[22]两种基于手工提取特征的关系抽取算法进行比较,其中 Mintz 代表文献[13]提出的传统的远程监督关系抽取模型,MIML 代表文献[22]提出的多实例多标签关系抽取模型. 图 6 是 PCNN + IATT 与两种算法的召回率-准确率对比图.

由图 6 可以看出:PCNN + IATT 模型在整个召回率范围内的准确率要远远高于 Mintz 模型和 MIML 模型,并且在召回率超过 0.1 之后,Mintz 模型和 MIML 模型

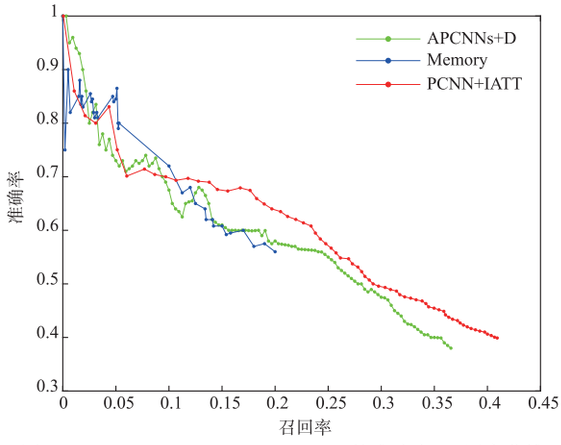

图5 APCNN+D, Memory, PCNN+IATT的准确率–召回率曲线

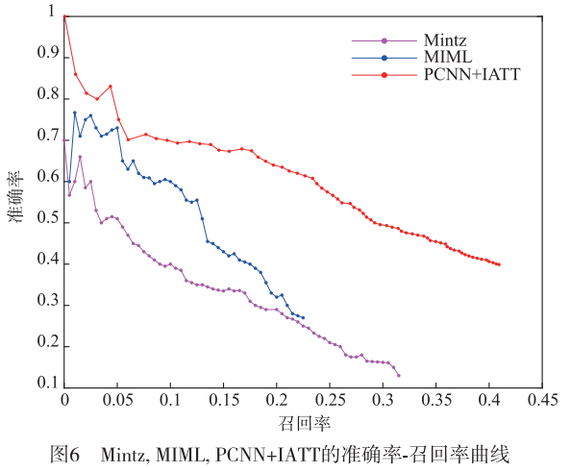

图6 Mintz, MIML, PCNN+IATT的准确率-召回率曲线

的准确率急剧下降,而PCNN + IATT模型的准确率下降速度平缓且在召回率达到0.4时仍保持了较可观的准确率.这说明人工设计的特征并不能很好地表达句子的语义信息并且使用NLP工具获取特征时带来的错误势必会损害模型的性能,而PCNN + IATT模型通过卷积神经网络自动学习到的句子特征则能充分地表达句子的语义信息.

3.3.4 实例比较

为了更加直观的对比各种算法的优劣,本文选取了一个标签为"/location/location/contains"的句子包,该包中包含9条句子,每条句子都包含实体对"Mexico,Acapulco".从表3中可以发现,第6、7、8条句子在语义上并没有体现"contains"关系,是噪声句子.本文采用文献[14,15]和本文相同的的权重计算方法计算每条句子的权重,并基于权重对句子进行排序,如表3所示.对于PCNN + ATT、APCNN + D以及Memory算法来说,其注意力模型就是求所有句子基于权重和的组合向量,即$0.238772 \times \boldsymbol{v}_1 + 0.264004 \times \boldsymbol{v}_2 + \cdots + 0.096322 \times \boldsymbol{v}_9$,这里$\boldsymbol{v}_i$表示第$i$条句子的特征向量($1 \leqslant i \leqslant 9$),虽然三条噪声句子的权重较低,但是都参与了模型训练.而本文则是按照式(7)组成多个组合向量$\boldsymbol{g}$,例如:$\boldsymbol{g}_1 = 0.264004 \times \boldsymbol{v}_2 / 0.264004$,$\boldsymbol{g}_2 = (0.264004 \times \boldsymbol{v}_2 + 0.238772 \times \boldsymbol{v}_1) / (0.264004 + 0.238772)$,然后基于softmax计算其在每个关系上的得分,然后采用在标注关系r上得分最高的组合句子向量来训练模型.通过实验发现,本例中PCNN + IATT算法选择$\boldsymbol{g}_5$作为组合向量来训练模型.可见$\boldsymbol{g}_5$包含5条权重最高的正实例句子,摒弃了所有噪声句子,虽然$\boldsymbol{g}_5$也丢掉了一条正实例,但是整体效果还是达到了最优,同时,也说明本文算法仍然存在改进的空间.

表3 实例比较

句子实例	权重	权重排序
1. Mr. gutierrez also made trips to Guadalajara and **Acapulco** in **Mexico**,to Amman in jordan and tel aviv in israel,as well as to less exotic destinations like tampa,san antonio and phoenix.	0.238772	2
2. Accompanied by my friend lynne scott,whom i've known since the long-ago days when i was briefly engaged to dan aykroyd -lrb- lynne was dating dan's brother,i arrive in **Acapulco**,**Mexico**,in early june to work on a cable movie,"romancing the bride."	0.264004	1
3. While the number of killings has gone down since president fox sent a battalion of federal officers to try to take back control of the city's streets,the violence has not ended but moved to other parts of **Mexico**,especially the central state of michoacan and the pacific coast resort of **Acapulco.**	0.036996	6
4. Leopoldo flores,a municipal judge in **Acapulco**,**Mexico**,officiated at casa de la laguna there.	0.174094	3
5. The last time rafael nadal and mariano puerta exchanged left hooks on clay was in february in the semifinals of a small tournament in **Acapulco**,**Mexico.**	0.139185	4
6. Unlike cancún or **Acapulco**,where american fast food joints and discos rule,san miguel de allende still feels like **Mexico.**	0.011007	9
7. The crisis erupted on dec. 2,when **Mexico's** top drug prosecutor,deputy attorney general josé luis santiago vasconcelos,acknowledged that eight agents in **Acapulco** had been arrested on charges that they had kidnapped hired assassins working for the gulf drug cartel and had handed them over to their rivals in the sinaloa cartel.	0.024685	7
8. The bill was approved as **Mexico** finds itself in the midst of a war between rival drug cartels that has claimed hundreds of lives,including dozens of police officers,particularly in the texas border town of nuevo laredo and along the pacific coast between **Acapulco** and zihuatanejo.	0.014933	8
9. In **Acapulco**,**Mexico**,i watched in dismay as tourists bought fruit at the marketplace and ate it,unpeeled.	0.096322	5

4 结论

为了最大限度地降低远程监督数据集中噪声数据的影响,本文提出一种基于改进注意力机制的卷积神经网络实体关系抽取模型. 该模型尽可能地从包含同一实体对的句子集中找出所有体现标签关系的正实例,构建组合句子特征向量,抛弃可能的噪声句子,从而最大程度地降低噪声句子对模型性能的影响同时又能充分利用句子集中所有正实例的语义信息. 实验证明,本文提出的模型在准确率上要优于几种经典的对比模型.

参考文献

[1] 陈宇,郑德权,赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报,2012,23(10):2572 – 2585.
CHEN Yu,ZHENG De-quan,ZHAO Tie-jun. Chinese relation extraction based on deep belief nets[J]. Journal of Software,2012,23(10):2572 – 2585. (in Chinese)

[2] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[A]. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics[C]. USA:Association for Computational Linguistics,2004. 415.

[3] Rozenfeld B, Feldman R. High-performance unsupervised relation extraction from large corpora[A]. Proceedings of the Sixth International Conference on Data Mining (ICDM′06)[C]. USA:IEEE,2006. 1032 – 1037.

[4] Gonzalez E, Turmo J. Unsupervised relation extraction by massive clustering[A]. Proceedings of the Ninth IEEE International Conference on Data Mining[C]. USA: IEEE, 2009. 782 – 787.

[5] Brin S. Extracting patterns and relations from the world wide web[A]. International Workshop on the World Wide Web and Databases [C]. Berlin: Springer, 1998. 172 – 183.

[6] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[A]. Proceedings of the Fifth ACM Conference on Digital Libraries[C]. USA: ACM,2000. 85 – 94.

[7] Liu X, Yu N. Multi-type web relation extraction based on bootstrapping[A]. International Conference on Information Engineering(ICIE)[C]. USA:IEEE,2010. 24 – 27.

[8] Chen J,Ji D,Tan C L,et al. Relation extraction using label propagation based semi-supervised learning[A]. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics[C]. USA: Association for Computational Linguistics,2006. 129 – 136.

[9] Moens M F. Information Extraction: Algorithms and Prospects in a Retrieval Context [M]. Germany: Springer, 2006. 315 – 317.

[10] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[A]. Proceedings of the ACL 2004 on Interactive Poster and Demonstration Dessions[C]. Spain: Association for Computational Linguistics,2004. 22.

[11] Jiang J,Zhai C X. A systematic exploration of the feature space for relation extraction[A]. Proceedings of NAACL HLT[C]. USA: Association for Computational Linguistics,2007. 113 – 120.

[12] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction[A]. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing[C]. USA: Association for Computational Linguistics,2005. 724 – 731.

[13] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[A]. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing [C]. Singapore: Association for Computational Linguistics,2009. 1003 – 1011.

[14] Zeng D,Liu K,Chen Y,et al. Distant supervision for relation extraction via piecewise convolutional neural networks[A]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing[C]. Portugal: Association for Computational Linguistics, 2015. 1753 – 1762.

[15] Lin Y, Shen S, Liu Z, et al. Neural Relation Extraction with Selective Attention over Instances[A]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1)[C]. Germany: Association for Computational Linguistics,2016. 2124 – 2133.

[16] Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao. Distant supervision for relation extraction with sentence-level attention and entity descriptions[A]. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)[C]. USA:AAAI,2017. 3060 – 3066.

[17] Xiaocheng Feng, Jiang Guo, Bing Qin, Ting Liu, Yongjie Liu. Effective deep memory networks for distant supervised relation extraction[A]. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17) [C]. Australia: IJCAI, 2017. 4003 – 4008.

[18] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011,12(Aug):2493 – 2537.

[19] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.
[20] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[A]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases[C]. Berlin: Springer, 2010. 148 – 163.
[21] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[A]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 1)[C]. USA: Association for Computational Linguistics, 2011. 541 – 550.
[22] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction[A]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning[C]. USA: Association for Computational Linguistics, 2012. 455 – 465.
[23] Zeiler M D. ADADELTA: an adaptive learning rate method[J]. arXiv preprint arXiv: 1212.5701, 2012.

作者简介

冯建周 男,1978 年 2 月出生于河北省沧州市. 现为燕山大学软件工程系副教授、硕士生导师. 主要研究方向为语义 web、知识图谱.
E-mail: fjzwxh@ysu.edu.cn

宋沙沙 女,1992 年 5 月出生于河北省沧州市. 现为燕山大学计算机科学与技术专业研究生. 主要研究方向为知识图谱.
E-mail: songshasha@he.chinamobile.com

王元卓 男,1978 年 7 月出生于黑龙江齐齐哈尔市. 现为中国科学院计算技术研究所科研处副处长、研究员、博士研究生导师. 主要研究方向为开放网络知识计算.
E-mail: wangyuanzhuo@ict.ac.cn

刘亚坤 男,1997 年 12 月出生于河北省唐山市. 现为燕山大学软件工程系本科生,主要研究方向为知识图谱.
E-mail: 2510278539@qq.com

武红颖 女,1995 年 8 月出生于河北衡水市. 现为燕山大学计算机技术专业研究生. 主要研究方向为知识图谱.
E-mail: mwuhongying@163.com

龚 昊 男,1997 年 4 月出生于河北省保定市. 现为燕山大学软件工程系本科生. 主要研究方向为知识图谱.
E-mail: aininot260@qq.com