

融合门控机制的远程监督关系抽取方法

李兴亚 陈钰枫[†] 徐金安 张玉洁

北京交通大学计算机与信息技术学院, 北京 100044; [†] 通信作者, E-mail: chenylf@bjtu.edu.cn

摘要 提出一种融合门控机制的远程监督关系抽取方法。首先在词级别上自动选择正相关特征, 过滤与关系标签无关的词级别噪声; 然后在门控机制内引入软标签的思想, 弱化硬标签对噪声过滤的影响; 最后结合句子级别的噪声过滤, 提升模型的整体性能。在公开数据集上的实验结果表明, 相对于句子级别噪声过滤方法, 所提方法的性能有显著提高。

关键词 关系抽取; 远程监督; 门控机制; 卷积神经网络

Distant Supervision for Relation Extraction with Gate Mechanism

LI Xingya, CHEN Yufeng[†], XU Jin'an, ZHANG Yujie

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;

[†] Corresponding author, E-mail: chenylf@bjtu.edu.cn

Abstract A piecewise convolutional neural network with gating mechanism is proposed, which would automatically filter positive correlation features at word-level. Moreover, the idea of soft-label is introduced to the gating mechanism to weaken the impact of hard labels on noise filtering. Combined with sentence-level noise filtering, the overall performance of the model is improved. The experimental results on the public dataset show that the proposed model has a significant improvement compared to the sentence-level noise filtering methods.

Key words relation extraction; distant supervision; gate mechanism; convolutional neural network

关系抽取任务旨在从自然语言文本中识别出两个实体间存在的语义关系, 是信息抽取任务的重要组成部分。关系抽取任务目前面临的挑战是模型训练需要大规模的标注语料, 而人工标注语料过于费时费力。Mintz等^[1]提出远程监督关系抽取方法, 使用知识库中的关系对训练语料进行启发式标注。例如, 给定一个知识库中的元组(也称关系事实), “Syracuse(实体1), contains(关系), Onondaga Lake(实体2)”, 远程监督方法会将包含上述两个命名实体的所有句子标注为 contains(包含)关系。

远程监督是一种快速有效的训练语料标注手段, 但存在严重的错误标注问题^[2]。例如, 在句子 “The Onondaga Nation is an 11-square-mile parcel in a valley south of Syracuse and about eight miles from

the southern end of Onondaga Laket”中, 实体 “Syracuse” 与实体 “Onondaga Lake” 之间不具有 contains 关系, 但会被视为正例。对此, Riedel等^[2]引入多示例学习, 将远程监督关系抽取建模为多示例单标签问题, 在包级别上预测标签。一些研究采用多示例多标签学习, 并使用概率图模型来选择句子^[3-5]。以上方法都严重依赖于NLP工具生成的特征质量, 深受错误传递问题的困扰。Zeng等^[6]基于 at-one-leaf 假设, 结合 MIL 和分段卷积神经网络 (piece-wise convolutional neural networks, PCNNs) 来选择最有可能是正例的句子。Lin等^[7]在 PCNNs 的基础上, 引入句子级别的注意力机制, 通过分配给句子不同的注意力权重来提取所有句子的有效特征。Ji等^[8]提出名为 APCNNs 的模型架构, 同样使用注意力机

国家自然科学基金(61976016, 61473294, 61370130, 61876198)、北京市自然科学基金(4172047)和科学技术部国际科技合作计划(K11F100010)资助

收稿日期: 2019-05-23; 修回日期: 2019-09-25

制, 并加入实体描述信息, 在此类方法中取得目前最好的效果。最近强化学习方法也用于远程监督关系抽取任务中^[9-10], 在模型训练之前使用相应方法对语料库中的句子进行筛选, 并取得较好的效果。

上述研究在远程监督关系抽取任务上取得很好的效果, 但仍存在两个问题。1) 句子中并非所有的单词都对关系标签的判断有贡献。例如, 句子“The cultural appreciation of photography in France was driven by literary theorists like Roland Barthes and Jean Baudrillard, not by artists”描述实体“Roland Barthes”与“France”之间的 nationality 关系, 但子句“not by artists”对关系“nationality”的判定作用较小, 从中提取的特征反而会影晌关系抽取模型的准确性。2) 训练过程中实体对的标签是不可变的(硬标签), 这在一定程度上扩大了远程监督错误标注问题带来的影响。

针对以上问题, 我们提出一种新的融合门控机制的 PCNNs 模型(gated PCNNs, GPCNNs), 利用门控机制^[11]对 PCNNs 提取的特征进行筛选, 使其保留更多与关系标签相关的特征。另外, 在门控机制中借鉴了软标签的思想^[12], 即标签信息会随着模型训练而改变。具体做法是, 将实体对之间的信息作为软标签来指导对错误标签相关特征的过滤, 从而减弱训练过程中错误标签带来的影响。最后, 结合句子级别的噪声过滤, 获得包含较多正相关特征的句子。在公开数据集上的测试结果表明, 我们提出的模型性能优于所有的基线系统, 验证了所提方法的有效性。

1 模型架构

给定一个包(bag, 包含相同实体对的句子集合) $B = \{s_1, s_2, \dots, s_n\}$ 和相应的两个实体, 本文模型在 B 上预测各个实体关系标签的概率。我们提出的融合门控机制的 PCNNs 模型整体架构如图 1 所示。

在图 1 中, 给定一个句子 s , 首先将句子中的单词转换为低维向量, 得到句子的矩阵表示, 之后使用门控 PCNNs 模块提取句子特征, 最后通过注意力机制得到最终的各个标签得分。

与传统的 PCNNs 模型相比, 我们对卷积层进行改进, 加入门控机制(图 1 中软标签监督的门控卷积层), 在此基础上引入软标签的思想, 使得模型可以更好地过滤词级别噪声。

1.1 向量表示层

词向量(word embedding)是单词的分布式表示, 目的是将单词映射到一个低维的向量空间。单词的向量表示通过查找预训练好的向量矩阵 V (或称查找表) 获得, 其中 $V \in \mathbb{R}^{|\mathcal{V}| \times d^w}$, $|\mathcal{V}|$ 为矩阵 V 大小, d^w 为词向量的维度。

我们采用位置特征(position feature)来确定实体对位置, 便于神经网络对单词与头实体(head entity) e_1 和尾实体(tail entity) e_2 之间的距离进行追踪。位置特征为当前单词到 e_1 和 e_2 相对距离的组合, 每个单词都有两个相对距离。图 2 为相对距离的例子, 词“literary”到 e_1 (France) 和 e_2 (Roland Barthes) 的相对距离分别为 4 和 -3。随机初始化两个位置向量矩阵 (PF_1 和 PF_2), 便可通过它们将相对距离映射为实

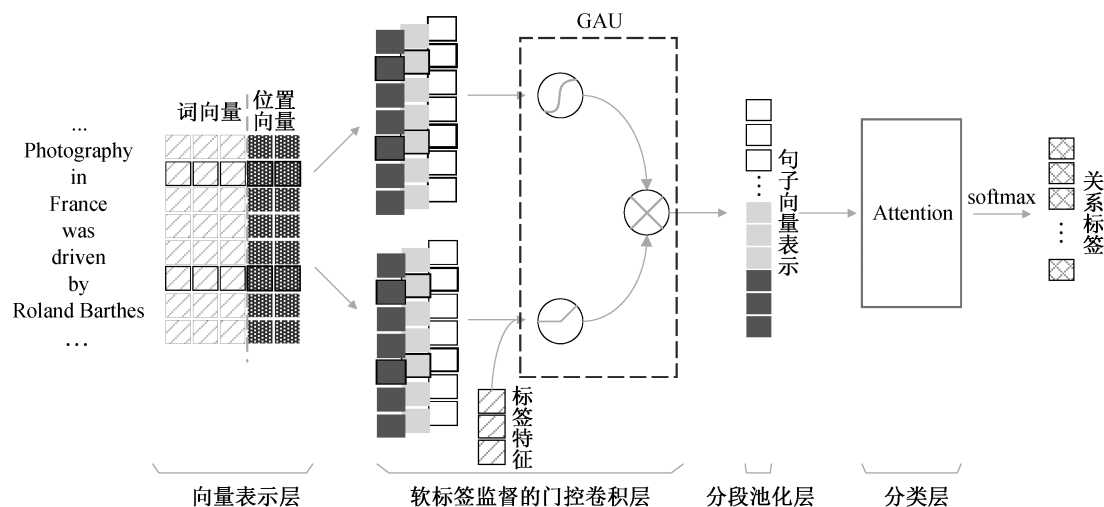


图 1 GPCNNs 整体架构
Fig. 1 Architecture of GPCNNs

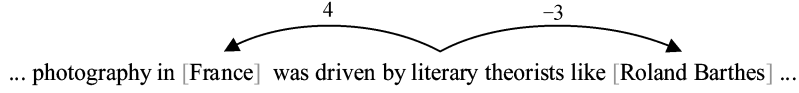


图2 相对距离的例子

Fig. 2 An example of relative distances

值向量, 即位置向量(position embedding)。

假设词向量的维度为 d^w , 且位置向量的维度为 d^p , 通过词向量和位置向量, 我们将句子 s 转换为一个矩阵 $\mathbf{S} \in \mathbb{R}^{|s| \times d}$, 其中 $|s|$ 为句子 s 的长度, $d = d^w + d^p$ 。然后将 \mathbf{S} 作为神经网络的输入(图1中的向量表示层)。

1.2 软标签监督的门控卷积层

卷积神经网络(convolutional neural network, CNN)可以有效地提取输入矩阵的所有局部特征, 并进行全局预测。并且CNN支持并行, 计算效率高。因此, 本文采用CNN来提取句子特征。

给定一个句子序列 $s = \{q_1, q_2, \dots, q_n\}$, 其中 q_i 为第 i 个单词的向量表示, $q_i \in \mathbb{R}^d$ 。我们将卷积核(window size)大小设置为 w (图1中的卷积层 $w = 3$), 用 $\mathbf{S}_{i,j}$ 表示 q_i 到 q_j 的拼接矩阵, 相应的卷积操作权重矩阵为 $\mathbf{w} \in \mathbb{R}^{w \times d}$ 。对 s 进行卷积操作后的结果为 $\mathbf{c} = \{c_1, c_2, \dots, c_{|s|-w+1}\}$:

$$c_j = f(\mathbf{w} \otimes \mathbf{S}_{(j-w+1):j} + b_c), \quad (1)$$

其中, $1 \leq j \leq |s|$, \otimes 为卷积运算, f 为非线性激活函数(如 \tanh 函数), $b_c \in \mathbb{R}$ 为偏置项。

考虑到词级别噪声对模型性能的影响, 本文采用门控机制来过滤词级别的噪声。基于 Gated Tanh Units(GTU)和 Gated Linear Units(GLU)的门控机制在语言模型中应用广泛^[13-14], 取得良好效果。综合考虑计算性能和有效性, 我们对基于GTU的门控机制进行改进, 将其命名为GAU(Gated Activation Units)模块, 对应的运算结构为

$$c_{\text{GAU}} = \tanh(\mathbf{w} \otimes \mathbf{S} + b_1) \times \text{relu}(\mathbf{v} \otimes \mathbf{S} + b_2), \quad (2)$$

其中, relu 函数会根据自身的输出来限制 \tanh 函数的输出, 达到过滤句内噪声的目的。

在基于硬标签的方法中, 实体对的关系标签(无论它们是否正确)在训练期间是不可变的, 这在一定程度上扩大了远程监督错误标注问题对模型性能产生的负面影响。本文在门控机制内引入软标签策略, 即训练时将标签替换为实体对之间的动态信息来指导特征过滤。

如图1所示, GAU模块连接两个卷积网络(一个为原始CNN, 另一个带有标签特征)。我们使用两个实体间的双线性变换结果 $l_{\text{relation}} = \mathbf{e}_1 \mathbf{W}_B \mathbf{e}_2$ (\mathbf{W}_B 为模型参数)作为实体对 $(\mathbf{e}_1, \mathbf{e}_2)$ 之间的关系标签, 指导模型在词级别上对不相关特征进行过滤。具体地, 通过以下方式获得特征 $c_{\text{GAU}, j}$:

$$m_j = \text{relu}(\mathbf{w}_{m,j} \otimes \mathbf{S}_{(j-w+1):j} + l_{\text{relation}} + b_m), \quad (3)$$

$$n_j = \tanh(\mathbf{w}_{n,j} \otimes \mathbf{S}_{(j-w+1):j} + b_n), \quad (4)$$

$$c_{\text{GAU}, j} = m_j \times n_j \circ \quad (5)$$

为了提取不同的特征, 通常需要使用不同的卷积核。因此本文采用 n 个卷积核, 每次卷积操作都需要两个系数矩阵, 即 $\mathbf{W} = \{(\mathbf{w}_{m,1}, \mathbf{w}_{n,1}), (\mathbf{w}_{m,2}, \mathbf{w}_{n,2}), \dots, (\mathbf{w}_{m,n}, \mathbf{w}_{n,n})\}$, 则卷积后得到的最终特征表示 c_{ij} 为

$$c_{ij} = c_{\text{GAU}, ij} = m_{i,j} \times n_{i,j}, \quad (6)$$

其中, $1 \leq i \leq n$, 且 $1 \leq j \leq |s| - w + 1$ 。经过以上运算, 卷积层的整体输出结果为 $\mathbf{C} = \{c_1, c_2, \dots, c_n\}$ 。

1.3 分段池化层

最大池化操作通常用于提取特征图(feature map)中最重要的特征, 但忽略了结构信息和细粒度信息。Zeng等^[6]提出一种新的池化机制, 根据给定的实体对将句子分为3部分, 分别进行最大池化。对于卷积层结果 $\mathbf{C} \in \mathbb{R}^{|s|-w+1}$, 可以划分为 $\mathbf{C} = \{c_{i,1}, c_{i,2}, c_{i,3}\}$, 则分段最大池化定义为 $p_{ij} = \max(c_{i,j})$, 其中 $1 \leq i \leq n$, $j = 1, 2, 3$ 。然后将所有的向量 $\mathbf{p}_i = [p_{i,1}, p_{i,2}, p_{i,3}]$ 进行拼接, 得到向量 $\mathbf{P} \in \mathbb{R}^{3n}$ (图1的分段池化层为 $n=3$ 时的情况)。最后通过计算得到句子 s 的特征向量 $\mathbf{b}_s = \tanh(\mathbf{P})$ 。

1.4 决策层

Lin等^[7]和Ji等^[8]分别使用双线性和非线性形式的注意力机制来弱化包中噪声句子对模型性能的影响。考虑到计算效率, 我们采用非线性形式的注意力机制。

我们同样使用 $l_{\text{relation}} = \mathbf{e}_1 \mathbf{W}_B \mathbf{e}_2$ 作为标签信息。对于第 i 个包 B_i 中的第 j 个句子的特征向量 \mathbf{b}_i^j , 其注意力权重 α^j 的计算如下:

$$\alpha^j = \frac{\exp(\omega^j)}{\sum_k \exp(\omega^k)}, \quad (7)$$

$$\omega^j = \mathbf{W}_a^T (\tanh[\mathbf{b}_i^j; I_{\text{relation}}]) + b_a, \quad (8)$$

则 B_i 的最终向量表示为 $\mathbf{r}_i = \sum_{j=1}^{|B_i|} \alpha^j \mathbf{b}_i^j$ 。

包的向量表示被送入 softmax 分类器, 来预测最终的关系标签, 并计算交叉熵目标函数:

$$J(\theta) = -\sum_{i=1}^T \log p(y_i | \mathbf{r}_i; \theta). \quad (9)$$

2 实验

2.1 数据集及评测指标

我们采用 Riedel 等^[2]公布的数据集来评测本文模型, 该数据集由 Freebase 中的关系事实(relation facts)和 New York Times (NYT)语料库对齐生成。我们将对齐后的 2005—2006 年的句子作为训练集, 2007 年的句子作为测试集。该数据集包含 NA 标签(指实体对之间不存在关系)在内的 53 种关系标签。训练数据包含 522611 个句子、281270 个实体对和 18252 个关系事实; 测试数据包含 172448 个句子、96678 个实体对以及 1950 个关系事实。

使用 held-out evaluation 方法评估本文模型。该方式提供一种近似的准确率(Precision)测量方法, 即分别从测试集和 Freebase 中发现关系事实, 并通过较来评估模型, 无需人工评估。

2.2 实验设置

使用预训练好的词向量可以提高模型收敛速度, 提升模型性能。因此, 本文使用 word2vec 工具^[15]在 NYT 语料库上训练词向量, 并使用出现频率在 100 以上的单词来构建词汇表。

在训练集上使用 3 折交叉验证来选择模型参数。卷积核大小取值范围为 {3, 5, 7}, 卷积核数量(feature maps)范围为 {50, 100, 200, 230}, batch size 取值范围为 {50, 100, 160}, 同时采用 dropout 策略和 Adadelta 优化算法来训练模型, 相应的学习率(learning rate)取值范围为 {0.001, 0.01, 0.1, 0.5}。最终获得各项超参数如下: 卷积核大小为 3; 卷积核数量为 230; batch size 为 50; 词向量维度和位置向量维度分别为 50 和 5; 学习率为 0.1; dropout 比例为 0.5。

2.3 模型整体性能评测

我们将 GPCNNs 模型与 3 个经典模型进行对比: PCNNs+MIL^[6], PCNNs+ATT^[7]以及 APCNNs^[8]。

PCNNs+MIL 选择包中得分最高的句子作为包的表示; PCNNs+ATT 以及 APCNNs 分别使用句子级别的双线性和非线性形式的注意力机制, 综合包内所有句子信息作为包的表示。另外, 我们将 GAU 添加到 PCNNs+MIL 模型中(PCNNs+MIL+GAU)来验证 GAU 的有效性; 把 GAU 替换成 GTU(PCNNs+GTU)来验证实体对的双线性变换, 作为软标签的有效性。图 3 展示本文模型和所有基线模型的总体 P - R 曲线, 表 1 显示 N 分别为 100, 200 和 300 时的 Precision@N 值。

通过图 3 和表 1 可以得到以下结论。

1) 所有加入 GAU 的模型预测结果都优于不加 GAU 的基线模型, 表现为在相同召回率的情况下, 加入 GAU 的模型取得更高的准确率。这说明与单纯的注意力机制相比, 门控机制结合注意力机制可以过滤掉更多的负相关特征, 验证了 GAU 模块的有效性。

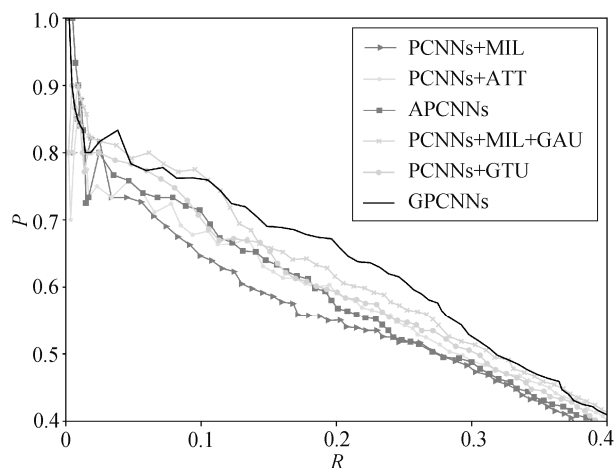


图 3 各个模型的总体 P - R 曲线

Fig. 3 Aggregate precision/recall curves for a variety of models

表 1 PCNN+MIL, PCNN+ATT, APCNNs 和 GPCNNs 的 Precision@N

Table 1 Precision@N of PCNN+MIL, PCNN+ATT, APCNNs and GPCNNs

模型	Precision@N/%			
	Top 100	Top 200	Top 300	平均
PCNNs+MIL	72.89	69.23	64.05	68.72
PCNNs+ATT	74.26	72.14	68.44	72.61
APCNNs	76.24	74.13	69.44	73.27
PCNNs+MIL+GAU	81.19	77.61	74.42	77.74
PCNNs+GTU	78.22	76.62	68.77	74.54
GPCNNs	83.16	77.60	74.44	78.13

2) GPCNNs 的整体 P - R 曲线相比于其他模型下降更为平缓, 曲线下覆盖面积最大; Precision@ N 值皆达到最高, 比基线模型评价提高约 4%。说明 GPCNNs 性能明显优于所有基线模型。

3) GPCNNs 取得比 PCNNs+GTU 更好的效果, 说明双线性变换可以有效地映射实体对间的关系特征, 结合门控机制可以有效地提取词级别的特征; 同时也证明软标签策略可以更加有效地筛选特征。

2.4 门控机制性能评测

为了更直观地验证门控机制的有效性, 我们通过以下两种方式去除注意力机制并进行验证: 1) 去除 GPCNNs 的注意力选择机制(PCNNs+GAU), 即在句子级别上进行关系抽取任务; 2) 将 GAU 单元应用于 CNNs(CNNs+GAU), 并将 CNNs 和 PCNNs 作为基线模型设置对比实验。实验结果见图 4。

从图 4 可以得到以下结论。

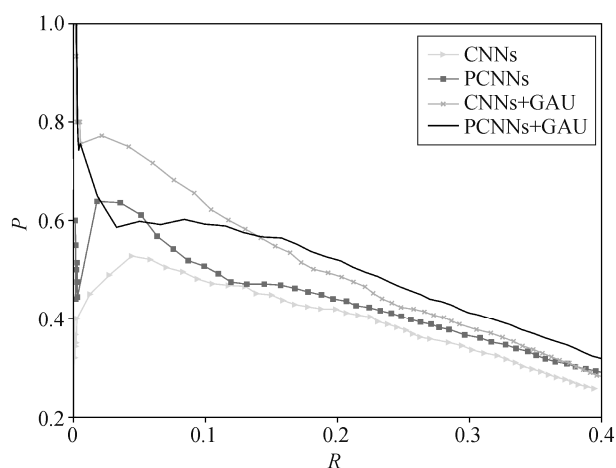


图 4 句子级别抽取中各个模型的总体 P - R 曲线
Fig. 4 Aggregate precision/recall curves for a variety of models on the sentence-level extraction

1) 相较于 PCNNs 和 CNNs 的性能, PCNNs+GAU 和 CNNs+GAU 都显著地提升, 说明在不同的池化机制下 GAU 模块都能有效地提高模型的整体性能, 从而验证了门控机制的有效性, 也反映出 GAU 模块的鲁棒性。

2) PCNNs+GAU 的整体曲线最平缓, 且 P - R 曲线下覆盖面积最大, 说明词级别噪声会对关系抽取模型的性能产生较大的影响, 验证了 GAU 模块的有效性, 同时表明 GAU 模块可以有效地过滤词级别噪声, 提高远程监督关系抽取模型的整体性能。

2.5 样例分析

为了证明词级别噪声对模型性能的影响以及本文模型在词级别噪声过滤上的有效性, 下面展示了测试过程中包中的句子注意力权重分配的一个例子。如表 2 所示, 该包共有 3 个句子, 均含有短语搭配“president emeritus of”, 该短语可以有效地表明实体“John Brademas”与“New York University”之间的“/business/person/company”关系, 因此 3 个句子均为正例。

可以看出, 表 2 的句子中除短语搭配“president emeritus of”外, 其他单词与实体对关系“/business/person/company”并不具有直接或间接的联系, 属于噪声。相对于其他句子, 第 1 个句子包含的噪声最多, 第 3 个句子次之, 第 2 个句子最少。由于缺乏有效的词级别噪声过滤机制, APCNNs 模型分配给第 2 句较高的权重(0.665), 第 1 句和第 3 句较少的权重(0.085 和 0.249), 而本文 GPCNNs 模型基本上不受噪声的影响, 分配给 3 个句子相近的权重(0.379, 0.313 和 0.308)。注意力权重的分配情况验证了本文模型可以有效地过滤词级别噪声, 并能更充分地利用包中的监督信息。

表 2 注意力权重 α 的例子
Table 2 An example of weight α

实体关系对	句子	α	
		APCNNs	GPCNNs
(John Brademas, New York University, /business/person/company)	1. Thirty-five years ago, President vetoed the legislation, refusing to encourage "the family-centered child rearing". "I don't think we've ever recovered from that veto message." said John Brademas, president emeritus of New York University, and, as a former Democratic congressman from Illinois, a sponsor of that legislation.	0.085	0.379
	2. An article on Jan.11 about a conference in New York misidentified the home state of John Brademas, president emeritus of New York University.	0.665	0.313
	3. Correction: January 25, 2006, Wednesday An article on Jan. 11 about a conference in New York misidentified the home state of John Brademas, president emeritus of New York University.	0.249	0.308

3 结语

针对词级别的噪声过滤问题, 本文提出一种新的融合门控过滤机制的分段池化卷积神经网络方法来进行远程监督关系抽取。门控机制可以有效地过滤词级别噪声, 结合句子级别的筛选机制, 显著地提升远程监督关系抽取模型的整体性能。另外, 我们在门控机制中引入软标签思想, 使用实体对之间的双线性变换来指导特征过滤, 提高词级别噪声过滤的精度。实验结果表明, 我们的模型优于所有的基线模型, 达到最好的效果。未来的研究中, 我们将结合强化学习, 引入先验知识, 探索进一步提高抽取性能的方法。

参考文献

[1] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data // 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 AFNLP. Singapore, 2009: 1003–1011

[2] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2010: 148–163

[3] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 1997, 89(1/2): 31–71

[4] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. Portland, 2011: 541–550

[5] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction // Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island Association for Computational Linguistics, 2012: 455–

465

[6] Zeng Daojian, Liu Kang, Chen Yubo, et al. Distant supervision for relation extraction via piecewise convolutional neural networks // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, 2015: 1753–1762

[7] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, 2016: 2124–2133

[8] Ji Guoliang, Liu Kang, He Shizhu, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions // Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, 2017: 1–7

[9] Feng Jun, Huang Minlie, Zhao Li, et al. Reinforcement learning for relation classification from noisy data [EB/OL]. (2018–08–24)[2019–04–20]. <https://arxiv.org/abs/1808.08013>

[10] Qin Pengda, Xu Weiran, Wang W Y. Robust distant supervision relation extraction via deep reinforcement learning [EB/OL]. (2018–05–24)[2019–04–20]. <https://arxiv.org/abs/1805.09927v1>

[11] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, 9(8): 1735–1780

[12] Wang Guanying, Zhang Wen, Wang Ruoxu, et al. Label-free distant supervision for relation extraction via knowledge graph embedding // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 2246–2255

[13] Kalchbrenner N, Espeholt L, Simonyan K, et al. Neural machine translation in linear time [EB/OL]. (2017–03–15)[2019–05–02]. <https://arxiv.org/abs/1610.10099>

[14] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks // Proceedings of the 34th International Conference on Machine Learning. Sydney, 2017: 933–941

[15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013–09–07)[2019–05–10]. <https://arxiv.org/abs/1301.3781>