

采用多尺度注意力机制的远程监督关系抽取*

蔡强^{1,2}, 郝佳云^{1,2}, 曹健^{1,2}, 李海生^{1,2}

(1. 北京工商大学, 计算机与信息工程学院, 北京市 100048;

2. 北京工商大学, 食品安全大数据技术北京市重点实验室, 北京市 100048)

摘要: 针对目前大多数关系抽取模型中局部特征以及全局特征利用不充分的缺点, 本文提出一种采用多尺度注意力机制的远程监督关系抽取模型。在词语层面, 通过在池化层构建权重矩阵来衡量词语与关系的相关程度从而捕捉句子中重要的语义特征; 在句子层面, 采用注意力机制将预测关系与句子进行相关性比较, 获得句子级别的重要信息。模型在 NYT 数据集上平均准确率达到 78%, 表明该模型能够有效地利用多尺度特征, 并且提高远程关系抽取任务的准确率。

关键词: 多尺度; 注意力机制; 远程监督模型; 关系抽取

中图分类号: TP391

文献标识码: A

Multi-level Attention-Based Distant Supervision for Relation Extraction

CAI Qiang^{1,2}, HAO Jiayun^{1,2}, CAO Jian^{1,2}, LI Haisheng^{1,2}

(1. School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, 100048, China; 2. Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing, 100048, China)

Abstract: Due to most of the existing relation extraction models couldn't make full use of the local and global feature, we proposed a distant supervised relation extraction model based on Multi-level attention mechanism. We employed an attention matrix in pooling layer to capture the word-level semantic feature which indicates the relevant relationship between input words and relations. Moreover, we adopted sentence-level attention mechanism to compare the relationship between sentences and predicted relations. Experimental results show that the mean accuracy of the proposed model is 78% in the NYT data set. The proposed model can effectively use multi-level feature and improves the accuracy of distant relation extraction task.

Key words: multi-level; attention mechanism; distant supervised model; relation extraction

1 引言

信息抽取技术在自然语言处理领域十分重要。关系抽取作为信息抽取的重要分支, 是用来识别文本中实体预先定义的语义关系^[1]。即对于实体对 e_1 和 e_2 , 二者之间相关关系可以形式化地表示为三元组形式 $\langle e_1, r, e_2 \rangle$, 其中 r 为关系描述类型。例如, 给定一个简单的包含实体关系的句子: “Bill Gates is the founder of the Microsoft.”, 其中实体对 “Bill Gates” 与 “Microsoft” 之间的关系为 “founder”。关系抽取技术已经被广泛应用于信息检索、基因疾病关系挖掘、知识图谱等重要领域。

近年来, 深度学习在大量自然语言处理任务中取得了超过了传统方法的性能, 因此大量算法采用了深度学习的方法进行特征提取以及关系抽取。2012 年, Socher^[2]提出使用递归神经网络来解决关系分类问题, 通过递归神经网络得到句子向量表示, 从而用于关系分类。之后, Zeng^[3]等人采用卷积神经网络结合词向量以及词语位置信息进行关系分类。虽然这些方法取得了较好效果, 但是在对模型进行训练时需要大量标注数据, 耗费了大量的人力物力。因此, 本文重点研究远程监督方法。

远程监督关系抽取首次由 Craven^[4]等人提出, 利用知识库信息来发现蛋白质与细胞/疾病/药物之间的关系。Mintz^[5]等人将知识与文本集进行对齐进行大规模关系抽取。但是, 错误标签引入了大量噪音,

* 收稿日期: 2017-06-10

定稿日期: 2017-07-25

基金项目: 北京市教委科研计划一般项目(SQKM201610011010); 北京市自然科学基金(4162019); 北京市科技计划课题(Z161100001616004)

作者简介: 蔡强, (1969—)男, 博士, 教授, 主要研究方向为计算机图形学、计算几何、科学可视化、智能信息处理; 郝佳云, (1993—)女, 硕士研究生, 主要研究方向为关系抽取, 知识图谱; 曹健, (1982—)男, 博士, 副教授, 主要研究方向为图像处理、模式识别等; 李海生, (1974—)博士, 教授, 主要研究方向为计算机图形学、科学可视化、三维模型检索等。

因此文献[6]提出了使用多示例学习的远程抽取策略，大规模减少了人工标注数据的工作，但是由于在句子编码时未充分利用句子中重要的语义信息，抽取结果准确率并不高。针对这一不足，一些算法进行关系抽取时采用了注意力机制的方式，用于丰富编码的语义信息并且减少编码过程中的噪声问题。

注意力机制曾在序列到序列任务中大放异彩，在对句子进行建模中取得了较好效果。因此，2016年，Lin^[7]等人提出了句子级别的注意力模型，用来降低远程监督关系抽取模型中错误标签带来的噪音问题。Zhou^[8]等人在采用长短期记忆模型（Long Short-Term Memory, LSTM）得到句子高层语义之后，使用注意力权重矩阵进行高层语义表示，提高了句子表示的准确性。但是这些方法在表征句子的局部及全局信息时仍有不足。

Yang^[9]等人曾将层次化的注意力机制应用到文本分类任务中，采用词语和句子层面的注意力模型对文本进行分类并且取得了不错的效果。而在关系抽取任务中词语和句子的特征向量表示对分类效果同样有着重要影响。在生成句子向量时，用于分类的关系对于句子编码的重要程度不同。例如，在句子“The burst has been caused by water hammer pressure.”中，关系“cause”对句子中词语的相关程度要强于关系“location”。因此考虑句子中各词语与关系之间的相关性影响着句子的向量表示。同时，在同一实体对对应的句子集合中，实体对在知识库中对应的关系对于不同句子的影响程度也不同。例如，实体对“Bill Gates”与“Microsoft”在知识库中对应的关系为“founder”，关系标签“founder”对于句子“Bill Gates is the founder of the Microsoft.”有较高的相关性，而对于句子“Bill Gates continues to serve on Microsoft’s Board as an advisor on key development projects.”相关性较低。所以，通过计算关系对不同句子的相关性一方面可以降低错误标签带来的噪音问题，另一方面可以获得不同句子中丰富的语义信息。

因此本文提出一种多尺度的注意力模型，采用注意力机制提取更加丰富的词语以及句子特征。模型使用双向门控循环单元（Bidirectional Gated Recurrent Unit, Bi-GRU）得到高层语义信息，在词语层面，通过在池化层采用权重矩阵来捕捉不同词语与关系之间的相关性；在句子层面，通过计算句子与知识库中对应的实体对之间预测关系的相关程度得到最终的句子向量表示。

2 多尺度注意力关系抽取模型

为了更好地利用句子语义信息，捕捉句子中较为重要的部分并且降低错误标签带来的噪音问题，本文结合词语层面以及句子层面的注意力机制，提出了多尺度注意力关系抽取模型，模型设计方式如图 1 所示：

- (1) 输入映射层：将词语与实体对之间位置向量作为神经网络模型输入。
- (2) Bi-GRU 层：采用双向 GRU 得到高层语义信息。

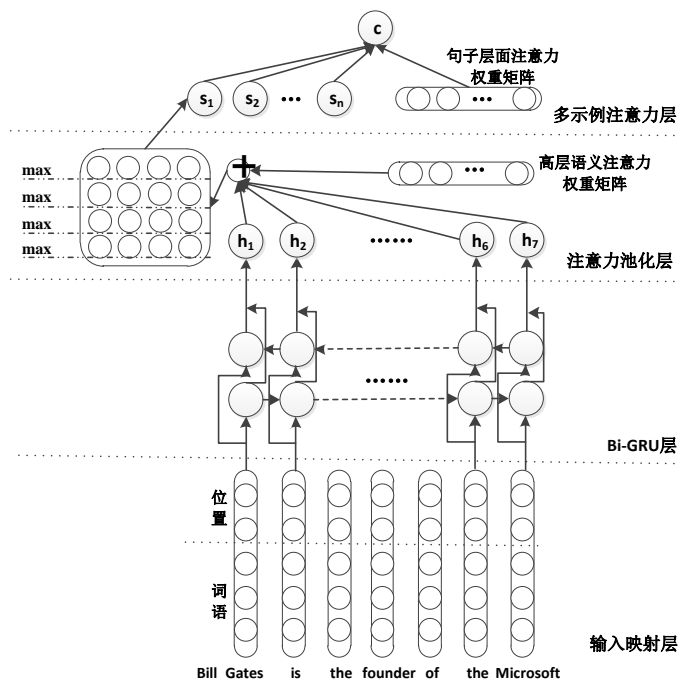


图 1 多注意机制关系抽取模型

(3) 注意力池化层：通过计算句子中词语与所有关系之间的相关程度，建立词语层面权重矩阵进行池化，并且将词语水平的向量合并成为句子水平向量。

(4) 多示例注意力层：计算实体对集合中句子向量和预测关系之间的相关程度，建立句子层面权重矩阵，得到最终的句子向量表示。

模型的具体设计细节如下：

2.1 输入映射层

为了捕捉词语的句法和语义信息，需要将输入句子中的词语映射为词向量。对于包含 m 个词语的句子 $s = \{w_1, w_2, \dots, w_m\}$ ，其中每个词语 w_i 均被表示为实值向量 \mathbf{w}_i 。

$$\mathbf{w}_i = \mathbf{W}^{word} \mathbf{v}^i \quad (1)$$

其中， $\mathbf{W}^{word} \in \mathbb{R}^{d_w \times |V|}$ 是由 word2vec 训练得到的向量矩阵， d_w 是词向量的维度， $|V|$ 是词典的大小， \mathbf{v}^i 是输入词语的词袋表示（one-hot 形式）。由此得到一个向量序列 $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$ 。

2.2 Bi-GRU 层

GRU 是 Chol^[10]等人提出的 LSTM 的一个变种，包含更新门和重置门 2 个门结构和 1 个隐藏状态。为了得到序列中过去和未来的上下文信息，本文采用双向 GRU 得到高层语义表示。根据文献[3]的假设，在关系抽取任务中，越靠近实体的词语包含抽取关系的信息越丰富，本文采用词向量以及词语位置向量映射作为双向 GRU 的输入，因此对于第 i 个词语的输入为：

$$x_i = \{w_i, p_{i,1}, p_{i,2}\} \quad (2)$$

其中， w_i 为第 i 个词语， $p_{i,1}$, $p_{i,2}$ 分别表示第 i 个词语与第一个实体和第二个实体间的位置距离。

在 GRU 中，新记忆 \hat{h}_i^c 是由过去的隐含状态 h_{i-1} 和新输入 x_i 共同得到：

$$\hat{h}_i^c = \tanh(l_i \circ U^{(n)} h_{i-1} + V^{(n)} x_i) \quad (3)$$

其中，重置信号 l_i 用于判定 h_{i-1} 对结果 \hat{h}_i^c 的重要程度，计算公式如下：

$$l_i = \sigma(V^{(m)} x_i + U^{(m)} h_{i-1}) \quad (4)$$

更新门 z_i 决定了过去隐含状态 h_{i-1} 向下一个状态传递的程度：

$$z_i = \sigma(V^{(s)} x_i + U^{(s)} h_{i-1}) \quad (5)$$

隐含状态 h_i 由过去隐含状态 h_{i-1} 和新的记忆 \hat{h}_i^c 产生：

$$h_i = (1 - z_i) \circ \hat{h}_i^c + z_i \circ h_{i-1} \quad (6)$$

其中， $V^{(n)}$, $U^{(n)}$, $V^{(m)}$, $U^{(m)}$, $V^{(s)}$, $U^{(s)}$ ，是在训练 GRU 时，学习得到的参数。

如图 1 所示，网络结构包括两个子网络，包括正向和反向的上下文信息，因此对于第 i 个词语的向量表示 h_i 由正向网络和反向网络输出的向量 \vec{h}_i 和 $\overset{\leftarrow}{h}_i$ 得到：

$$h_i = [\vec{h}_i \oplus \overset{\leftarrow}{h}_i] \quad (7)$$

2.3 注意力池化层

对于关系抽取任务，用于分类的关系集合对于句子中词语的重要程度不同。因此，本文采用词语层面的注意力权重矩阵捕捉句子中与目标关系更加密切的信息。不同于传统的池化操作，为了得到与分类任务更相关的特征，本文采用注意力机制的池化操作。将通过双向 GRU 层得到的句子向量与注意力权重矩阵相乘，之后采用最大池化的操作获得最显著的特征表示，从而将词向量转化为句子向量。

在 2.2 小节得到的句子 H ($H \in \mathbb{R}^{d \times t}$ ， d 为经过双向 GRU 层后得到的表示单个词语向量的维度， t 为句子的长度) 表示为 $[h_1, h_2, \dots, h_t]$ 。所有关系组成的集合为 \mathbf{Y} ($\mathbf{Y} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_l\}$ ， \mathbf{r} 是关系的向量表示， l 是关系的数量)，如图 2 所示，通过计算句子向量和关系向量的内积，得到句子及关系相关度权重矩阵 $\mathbf{U}^{(0)}$ ：

$$\mathbf{U}^{(0)} = \text{softmax}(\mathbf{H}\mathbf{V}^{(0)}\mathbf{Y}) \quad (8)$$

其中，参数矩阵 $\mathbf{V}^{(0)}$ ($\mathbf{V}^{(0)} \in \mathbf{R}^{d \times l}$) 是在训练过程中更新得到。

通过将经双向 GRU 层得到的句子向量 \mathbf{H} 与权重矩阵相乘，从而突出词语层面的重要部分。之后，采用文献[11]的策略，采用最大化的策略选择最显著的特征。因此，句子表示为：

$$\mathbf{s} = \text{maxpool}(\mathbf{H}\mathbf{U}^{(0)}) \quad (9)$$

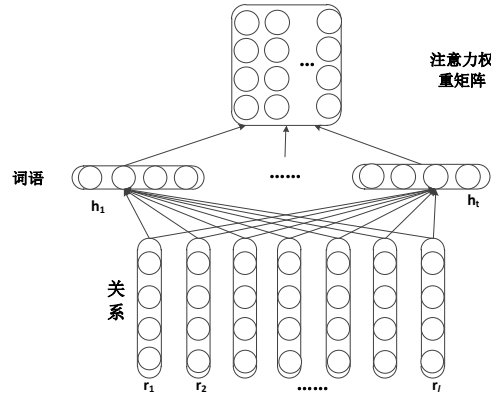


图 2 注意力池化层权重矩阵

2.4 多实例注意力层

在传统的远程监督抽取关系任务中，不可避免会引进错误标签，从而为关系抽取带来噪音。针对这一问题，本文采用多实例建模^[7]的方式，对于实体对，考虑实体与预测关系之间的相关程度，建立注意力矩阵，降低噪音对正确关系的影响，并且充分利用这些句子中的语义信息得到最终句子向量表示。

对于包含相同实体对的句子集合 S ，假定其中包含句子的数目为 n ，即 $S = \{s_1, s_2, \dots, s_n\}$ 。由 2.3 小节得到的集合 S 中句子向量可以表示为 $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ ，为了计算输入句子 s_i 与关系 r 之间的相关程度，通过计算句子集中句子向量与知识库中实体对对应关系向量的内积，得到注意力矩阵。

权重矩阵的计算公式如下：

$$\alpha_i = \text{softmax}(\mathbf{s}_i \mathbf{A} \mathbf{r}) \quad (10)$$

其中， \mathbf{A} ($\mathbf{A} \in \mathbf{R}^{d \times d}$) 为加权对角矩阵， \mathbf{r} 是实体对在知识库中对应的预测关系 r 的向量表示，由于关系向量在测试过程中是未知的，因此，其在训练过程中为实体对在知识库中对应的预测关系，在测试过程中通过随机初始化得到。

为了使与关系向量更为相关的句子被赋予较高权重，因此，将实体对对应的句子表示为：

$$\mathbf{c} = \sum_i \alpha_i \mathbf{s}_i \quad (11)$$

最后，采用 softmax 分类器从所有关系集 \mathbf{Y} 中预测实体对的关系标签 \hat{y} ：

$$\hat{p}(y | H) = \text{softmax}(\mathbf{Y}\mathbf{c} + \mathbf{b}) \quad (12a)$$

$$\hat{y} = \underset{y}{\text{argmax}} \hat{p}(y | H) \quad (12b)$$

其中， \mathbf{b} 为偏置向量。

2.5 训练和优化策略

本文采用交叉熵代价函数作为目标函数，定义如下：

$$J(\theta) = \sum_{i=1}^T \log p(r_i | H_i, \theta) \quad (13)$$

其中， θ 表示模型中所有的参数， T 代表句子集合数，本文使用 Adam 优化器进行参数更新。

为了防止模型过拟合，本文采用 Dropout 进行正则化约束。Dropout 最先是由 Hinton^[12]等人提出，在每次前向传播时，随机地丢弃一些隐层节点特征，即权值更新不依赖于固定的节点共同作用。本文在双向 GRU 层采用 Dropout。

另外，本文采用了 L2 正则化，在迭代时乘以一个小于 1 的因子 λ ，用于减小参数 θ 的值。正则化

操作降低了数据偏移对结果的影响，增强了模型的抗扰动性，避免了过拟合现象。

3 实验结果及分析

3.1 数据集及评价准则

为了评估多尺度注意力关系抽取模型，本文采用 2010 年由 Riedel^[13]等人提出的数据集。该数据集是将知识库 Freebase 和文本集 New York Times 通过启发式的匹配对应生成的，并被广泛应用于远程抽取任务中。具体地，本文采用的是 2005-2006 年的句子作为训练示例，2007 年的句子作为测试示例。数据集中包含 53 种关系（包含“NA”，表示实体对之间没有关系），其中训练集中包含实体对数目为 281,270，测试集中包含实体对个数为 96,678。

为了评价本文的方法是否有效，本文采用平均准确率（P@N）、准确率-召回率（PR）曲线来进行评价。通过对比前 N 项准确率以及 PR 曲线下的面积来评估算法的好坏。

3.2 参数设置

在实验过程中，采用交叉验证的方式进行模型调优，验证集从训练集中随机抽样获取。参数设置的过程参考文献[7]中的经验值，句子向量的维度取值范围为{50,60,...,300}；关系向量的维度与句子向量一致；学习率的取值范围为{0.01,0.001,0.0001}；批大小的取值范围为{50,100,150,200}。经过实验，本文采取的参数设置如表 1 所示。

表 1: 参数设置

词向量维度 d_w	句子向量维度 k_s	位置向量维度 k_p	关系向量维度 k_r	学习率 η	批大小 B	Dropout 参数 D	L2 正则项超参 λ
50	230	5	230	0.001	50	0.5	0.0001

3.3 实验验证

为了验证多尺度注意力模型对关系抽取性能的提高，本文将单注意力机制以及采用多尺度注意力机制对模型影响的效果进行了对比，结果如图 3 和表 2 所示，其中 sentence 表示仅采用句子层面的注意力模型，all 表示本文提出的多尺度注意力模型；表 2 是这两种方法前 100, 200, 300 的准确率以及平均准确率。从图表中可以看出，相较于采用了单个注意力机制的模型，结合多种尺度的注意力模型提高了关系抽取的准确性。

另外，本文与五种已经公开发表的方法进行了对比，如图 4 所示。Mintz 是由 Mintz^[5]等人提出，采用全部示例来抽取特征，MultiR^[14]采用了多示例学习的方法，MIMLRE^[15]采用多示例多标签的方法，CNN+ATT 与 PCNN+ATT 是由 Lin^[7]等人提出，采用 Zeng 等人在文献[3]和[6]中的工作，增加句子注意力机制的方法所得到的模型。图 4 表明，相较于其它模型，我们提出的模型有相对较高的准确率和召回率。另外，本文采用 GRU 获得句子的向量表示，相较于 CNN 的方式更能表征句子中的上下文信息，但同时对于局部特征的表征要弱于 CNN，因此如图 4 所示，在获得较高召回率的同时，由于引入了更多的噪音，对句子的向量表示影响较大，所以本文方法在准确率上要弱于采用 CNN 的方法。未来我们也将尝试将 CNN 与 GRU 进行结合用于表征句子向量，以获得更加丰富的特征。

表 2: 句子层面注意力模型与多尺度注意力模型准确率表

注意力模型	P@100	P@200	P@300	Mean
Sentence	0.76	0.73	0.7	0.73
all	0.84	0.77	0.73	0.78

4 结束语

在本文中，我们提出了一种采用多尺度注意力机制的模型。在词语层面以及句子层面均采用了注意力机制。充分利用了关系对于句子中词语的影响，并且考虑到了同一实体对在句子集合中，预测关系对句

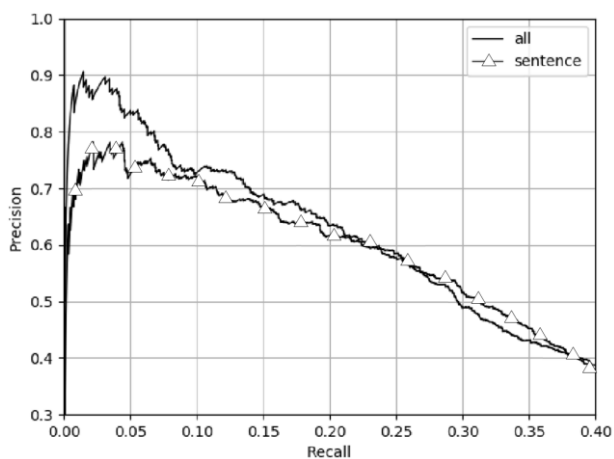


图3 句子层面注意力模型与多尺度注意力模型对比曲线

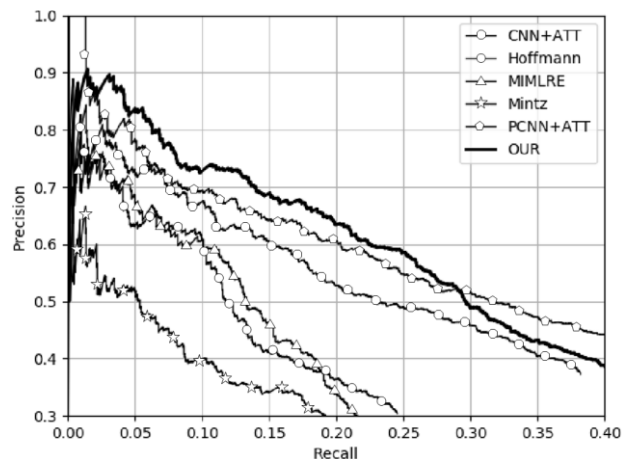


图4 本文方法与其他五种方法对比曲线

子编码的影响。实验表明，本文提出的模型适用于远程实体关系抽取任务。未来工作将尝试采用多类模型表征句子向量；并且在句子注意力机制方面，探索不同的方式解决多示例带来的噪音问题。

参考文献：

- [1] Li J, Zhang Z, Li X, et al. Kernel-based learning for biomedical relation extraction[J]. Journal of the Association for Information Science and Technology, 2008, 59(5):756 - 769.
- [2] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]// EMNLP-CoNLL 2012. Korea: [s. n.], 2012:1201-1211.
- [3] Zeng D, Liu K, Lai S, et al. Relation Classification via Convolutional Deep Neural Network[C]// COLING 2014. Ireland: [s. n.], 2014: 2335-2344.
- [4] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources[C]// ISMB. 1999. Heidelberg: [s. n.], 1999: 77-86.
- [5] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]// ACL-IJCNLP 2009. Singapore: [s. n.], 2009:1003- 1011.
- [6] Zeng D, Liu K, Chen Y, et al. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks[C]// EMNLP 2015. Lisbon: [s. n.]. 2015: 1753-1762.
- [7] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]// ACL 2016. Berlin: [s. n.]. 2016: 2124-2133.
- [8] Zhou P, Shi W, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// ACL 2016. Berlin: [s. n.]. 2016:207-212.
- [9] Yang Z, Yang D, Dyer C, et al. Hierarchical Attention Networks for Document Classification[C]// NAACL 2016. San Diego:[s. n.]. 2016:1480-1489.
- [10] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
- [11] Santos C, Tan M, Xiang B, et al. Attentive Pooling Networks[J]. arXiv preprint arXiv:1602.03609, 2016.
- [12] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4):212-223.
- [13] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[J]. Machine learning and knowledge discovery in databases, 2010: 148-163.
- [14] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]// ACL HLT 2011. Portland, Oregon, USA: DBLP, 2011:541-550.
- [15] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction[C]// EMNLP-CoNLL 2012. Korea: [s. n.]. 2012:455-465.

作者联系方式:

姓名: 郝佳云 地址: 北京市海淀区北京工商大学东校区 邮编: 100048 电话:18810578027 电子邮箱:
haojiayun_happy@163.com