文章编号:1007-5321(2020)05-0071-06 **DOI**:10.13190/j.jbupt.2020-071

基于高速多核网络的远监督关系抽取方法

李 威[1], 陈曙东[1,2], 欧阳小叶[2], 杜 蓉[2], 王 荣[3]
(1. 中国科学院大学 微电子学院, 北京 100049; 2. 中国科学院 微电子研究所, 北京 100029;
3. 北京跟踪与通信技术研究所 空间目标测量重点实验室, 北京 100094)

摘要: 远监督作为一种能够快速大量产生标注数据的技术,在关系抽取任务中的应用愈加广泛,但仍存在文本特征提取不足、包内噪声过多等问题. 对此,提出了一种基于高速多核网络的远监督关系抽取方法. 首先通过高速网络和多核卷积对句子特征进行深层提取;然后采用包内注意力机制提高包内正确标注的句子权重,降低包内噪声,实现包级向量化;使用包间注意力机制降低包间噪声,得到组级向量化;最后,将组作为训练样本训练分类器,实现关系抽取. 实验结果表明,该方法比现有方法具有更好的关系抽取性能.

关 键 词: 关系抽取; 远监督; 注意力机制; 神经网络; 高速多核网络模型
中图分类号: TP391 **文献标志码**: A

Distant Supervision Relation Extraction Method Based on Highway Multi-Kernel Network

LI Wei[1], CHEN Shu-dong[1,2], OUYANG Xiao-ye[2], DU Rong[2], WANG Rong[3]
(1. School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100049, China;
2. Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China;
3. Key Laboratory of Space Object Measurement, Department, Beijing Institute of Tracking and Telecommunications Technology, Beijing 100094, China)

Abstract: As a technology that can quickly generate large amounts of labeled data, the distant supervision is increasingly used in relation extraction. However, there are still problems such as insufficient text feature extraction and noise in the bag. A distant supervision relation extraction method based on highway multi-kernel network is proposed to solve these questions. Firstly, the feature of sentences are deeply extracted by highway network and multi-kernel convolution; and then the intra-bag attention mechanism is used to improve the sentence weight of the correct annotation in bag and reduce the intra-bag noise to obtain the bag's embedding. Subsequently, the inter-bag attention mechanism is used to reduce the inter-bag noise for each group of bags with the same relation to obtain the group's embedding. Finally, groups are used as training samples to train the classifier to achieve relation extraction. Experiment shows that this method has better relation extraction performance than existing methods.

Key words: relation extraction; distant supervision; attention mechanism; neural network; highway multi-kernel network

关系抽取是知识图谱领域的一个重要子任务, 目的是提取文本中实体之间的语义关系. 文本中表

收稿日期: 2020-06-20
基金项目: 中国科学院战略性先导科技专项(C 类)(XDC02070600)
作者简介: 李 威(1995—), 男, 硕士生.
通信作者: 陈曙东(1977—), 女, 研究员, 博士生导师, E-mail: chenshudong@ime.ac.cn.

达事实信息的名词被称作实体. 例如时间、组织机构、人物等. 在知识图谱的重要子任务关系抽取中,给定文本及文本所包含实体情况下如何对实体间的关系做出准确的判断至关重要. 除了实体之外,现存数据库的数据和网络文本数据还包含大量的关系信息,如"雷军是小米科技有限责任公司的创始人"中就包含了雷军、创始人、小米科技有限责任公司这三元组的关系信息. 目前一般将实体关系鉴别和实体关系分类合并为实体关系分类任务,即给出一个句子和句子中所包含的实体对,用训练好的分类器确定目标实体对之间的正确关系.

远监督学习将已有的大型知识库对应到非结构化文本数据中,如新闻文本、学术期刊等,从而自动生成大量的可用于关系抽取模型训练的训练数据. 远监督学习在关系抽取应用中有一个著名的假设:如果知识库中2个实体间具有某种关系,那么文本数据中所有提及这2个实体的句子都表示此种关系. 由于知识库中实体间的关系和文本中表达的实体间的关系存在差异,所以此假设会产生大量错误标注的训练数据,这些标注错误的训练数据为噪声训练数据. Riedel 等[1]证明将 Freebase 和 New York Times 数据集对齐产生的标注数据中正确标注的数据仅占总体的70%左右,余下的均为噪声训练数据,噪声训练数据会对关系抽取模型的训练效果产生负面影响.

针对远监督关系抽取任务中特征提取不足的问题,提出了高速多核网络模型(HMKN, highway multi-kernel network),创新点在于:① 提出的模型是一种性能优异的组级别远监督关系抽取算法;② 在 HMKN 模型中利用高速网络和多核卷积结合的方式对文本数据进行特征提取,使得特征和梯度的传递更加有效,提高了特征提取效果;③ HMKN 模型在 NYT 数据集上的平均分类准确率比远监督关系抽取模型提高了1.5%.

1 相关工作

为了解决标注数据数量不足的问题,Mintz 等[2]提出了远监督的概念:使用 Freebase 作为知识库进行远监督,将原始文本自动对应到知识库,以生成实体对的关系标签,并应用于知识图谱关系抽取的研究中. 这种远监督的假设会带来严重的错误标注问题,为了减弱错误标注对标注数据质量的负面影响,Riedel 等[1]提出了一个新的假设 EALO(expressed-at-least-one):若知识库中2个命名实体之间有某种关系,那么文本数据中必将存在"包含这2个实体并表达同样的关系"这样的句子. 这种假设相较于之前的假设更宽松,并且更加贴合实际情况,但同时也比前者更加复杂. 例如,远监督根据关系类型"place_of_birth"自动标注产生的2个句子,分别是A:"Barack Obama was born in the United States"和B:"Barack Obama was the 44th president of the United States". Riedel 将这些根据同种关系类别远监督标注产生的句子归纳进一个包中,并根据包内句子的表示导出包的表示. 其中句子 A 为标注正确的示例,而句子 B 并没有反映"place_of_birth"的关系为标注错误的数据,即噪声数据.

Ding 等[3]提出的非对称卷积网络(ACnet, asymmetric convolution net)使用多核非对称卷积的方法,替换目前 CNN 架构中常用的方形卷积核. 冀等[4]采用深度学习全连接神经网络对文本和评分函数进行特征学习,提高了深度学习模型的准确度和弹性. 在远监督关系抽取方面, Zeng 等[5]提出了切分卷积神经网络,卷积层的输出按照2个目标实体所处的位置划分为3部分,对其分别进行最大池化操作. Lin 等[6]将注意力机制加入关系抽取远监督任务的多示例学习中,对包内的噪声数据进行建模,在计算包的表示的过程中尽量减弱噪声数据对计算结果的影响,其实验结果证明了注意力机制在消除噪声数据负面影响方面的有效性. Ye 等[7]首次提出在远监督关系抽取任务中引入组的概念,将关系标签相同的包整合到一个组内,通过组内的注意力机制缓解了普遍存在的包噪声问题.

2 高速多核网络模型

HMKN 模型的框架如图1所示,模型由句子嵌入模块、包嵌入模块和组嵌入模块3部分组成. 首先,通过词嵌入的方法将数据集中的文本数据转化为向量的形式并获得句子的表示向量,再将向量作为句子嵌入模块的输入,通过切分卷积神经网络(PCNN, piecewise convolutional neural networks)结合高速多核网络的方法对句子表示向量进行特征提取,并将提取结果作为之后模块的输入;然后,在包嵌入模块中,将所有包含相同实体对句子的向量放入一个包中,并通过注意力机制获得包的表示向量;随后,在组嵌入模块中将所有被分类为同一种关系类型包的表示向量放进一个组中,并通过自注意力

机制获得组的表示向量. 令 $\boldsymbol{S}$ 表示经过特征提取后的句子向量，$\boldsymbol{B}$ 表示包的表示向量，$\boldsymbol{G}$ 表示组的表示向量.

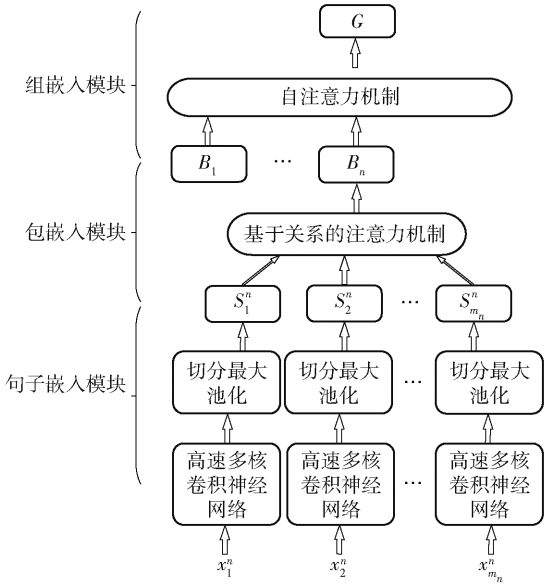

图 1　HMKN 高速多核网络模型

2.1　句子嵌入模块

数据集中每个句子的每个单词首先被影射为一个 d_w 维的词表示向量，并将 Zeng 等[5]提出的位置特征(PFs，position features)应用到 HMKN 中. 对于句子中的每个单词，PFs 描述了当前单词和头尾实体单词的相对距离，并生成 2 个 d_p 维的向量，丰富了目标实体语义特征. 最终将这 3 个向量接连成为一个 d_w+2d_p 维向量作为单词的表示向量.

提出了一种句子特征提取模块高速多核卷积神经网络，其结构如图 2 所示.

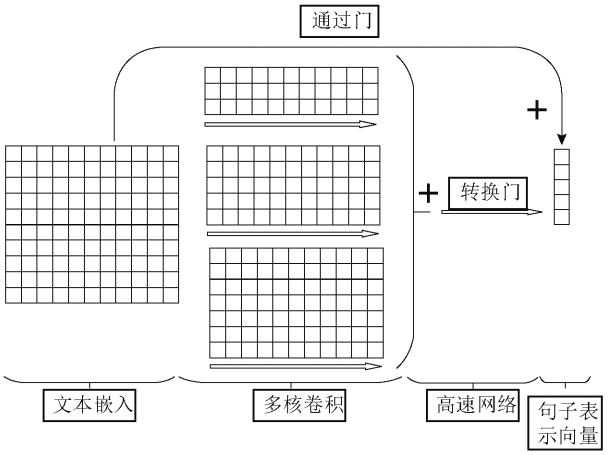

图 2　高速多核卷积神经网络

分别通过 3 个不同尺寸的卷积核对单词表示向量组成的矩阵 $\boldsymbol{W}_j^i \in R^{l_{ij}\times(d_w+2d_p)}$ 进行特征提取，得到卷积后的矩阵 $\boldsymbol{W}_{u1j}^i, \boldsymbol{W}_{u2j}^i, \boldsymbol{W}_{u3j}^i$. 将这 3 个输出的矩阵相加，有

$$\boldsymbol{W}_{u1j}^i + \boldsymbol{W}_{u2j}^i + \boldsymbol{W}_{u3j}^i = \boldsymbol{W}_{uj}^i$$

其中：i 为包中组的序号，j 为组中句子的序号. 将 $\boldsymbol{W}_{uj}^i$ 作为单词表示向量非线性变换的结果. 通过门控机制即 2 个非线性转换层、转换门和通过门将单词向量特征的线性变换结果和非线性变换结果加权相加，并将相加结果作为句子的表示向量.

根据句子中 2 个目标实体的位置，将高速网络的输出 Y 划分为 3 个部分，并使用 3 个最大池化层分别对这 3 个部分进行特征提取，得到句子的表示向量 $\boldsymbol{s}_j^i \in R^{3d_c}$. 令一个包中所有句子的表示向量为矩阵 $\boldsymbol{S}^i \in R^{m_i\times 3d_c}$，关系表示向量为矩阵 $\boldsymbol{R} \in R^{h\times 3d_c}$，其中 h 为关系类型数量.

2.2　包嵌入模块

将表示向量作为查询向量加入注意力机制的计算中，计算包中的句子与每种关系的匹配得分为

$$e_{kj}^i = \boldsymbol{r}_k \boldsymbol{s}_j^{i\mathrm{T}}$$

其中：$\boldsymbol{r}_k$ 为关系嵌入矩阵的第 k 行向量，$\boldsymbol{s}_j^i$ 为组中第 i 个包中第 j 个句子的表示向量. 为了最终通过注意力机制计算包的表示，需要赋予包内不同句子不同的权重，包内句子的权重定义为

$$\alpha_{kj}^i = \frac{\exp(e_{kj}^i)}{\sum_{j'=1}^{m_i}\exp(e_{kj'}^i)}$$

权重 α_{kj}^i 反映了包中第 j 个句子和第 k 个关系的相关程度，每个组中第 i 个包的表示向量为

$$\boldsymbol{b}_k^i = \sum_{j=1}^{m_i}\alpha_{kj}^i \boldsymbol{s}_j^i \tag{1}$$

所有包的表示向量 $\boldsymbol{b}^i$ 组成一个矩阵 $\boldsymbol{B}^i \in R^{h\times 3d_c}$，其中 $k\in\{1,2,\cdots,h\}$ 表示关系索引.

2.3　组嵌入模块

完成了包嵌入表示后，把表达同一种关系的所有包放进一个组里，对包和组内的其他所有的包分别做相似度计算，并将相似度得分加入包的权重计算中. 在注意力计算的过程中，使用自注意力算法，即组内不同包的注意力权重是按照使用每个组所包含所有包自身的表示向量计算的，即首先计算每个包自身的自身匹配得分.

$$\chi_{ik} = \sum_{i'=1,\cdots,n,i'\neq i}\mathrm{Sim}(\boldsymbol{b}_k^i, \boldsymbol{b}_k^{i'})$$

$$\mathrm{Sim}(\boldsymbol{b}_k^i, \boldsymbol{b}_k^{i'}) = \boldsymbol{b}_k^i \boldsymbol{b}_k^{i'\mathrm{T}}$$

每个χ_{ik}反映了第i个包和第k种关系之间的置信度,组内包的权重定义为

$$\beta_{ik} = \frac{\exp(\chi_{ik})}{\sum_{i'=1}^{m_i}\exp(\chi_{i'k})}$$

由此可得最终组的表示向量为

$$\boldsymbol{g}_k = \sum_{i=1}^{n}\beta_{ik}\boldsymbol{b}_k^i \tag{2}$$

其中:所有的$\boldsymbol{g}_k$组成一个矩阵$\boldsymbol{G}\in R^{h\times 3d_c}$,$\boldsymbol{g}_k$为$\boldsymbol{G}\in R^{h\times 3d_c}$的第$k$行.

2.4 目标函数设置

在模型的训练阶段,将每一个组作为模型训练的基本单位. 通过关系向量矩阵和组的表示向量计算出某个组分类到某个类别标签的得分为

$$o_k = \boldsymbol{r}_k\boldsymbol{g}_k^{\mathrm{T}} + d_k$$

其中:d_k为偏置项,最终通过softmax分类器对得分函数进行处理得到某个组分类为某个关系的概率大小,有

$$p(k|g) = \frac{\exp(o_k)}{\sum_{k'=1}^{h}\exp(o_{k'})}$$

整个模型的目标函数设置为

$$J(\theta) = -\sum_{(g,k)\in T}\ln p(k|g;\theta) \tag{3}$$

其中:T为所有训练样本的集合,θ为模型的参数集合.

3 实验

3.1 数据集和评估指标

采用NYT(New York Times)数据集进行实验验证. 数据集中包含52种普通关系和一个“无关系”(没有关系). 评估指标包括:P@N值、精确率召回率曲线和AUC值.

3.2 参数设置

在Pytorch 0.4.1中实现了所有的神经网络模型,将随机梯度下降法作为优化器,在多核卷积阶段,卷积神经网络使用3个不同尺寸的卷积核:(3,60),(5,60),(7,60),每种尺寸使用230个卷积核,模型的batchsize设置为20,group size设置为5. 在dropout阶段,训练时将隐藏单元随机为0的概率设置为0.5,初始学习率设置为0.1.

3.3 实验结果和分析

表1所示为HMKN模型及其变体在这3个测试集上的测试结果和对比方法下的实验结果. 对比方法包括:PCNN + ATT[6]是远监督关系抽取最好的方法之一,通过使用句子级别的注意力机制来选取对关系提取有影响的句子;PCNN + ATT + soft-label是Liu等[8]对PCNN + ATT的改进结果,BGRU + 3ATT是李等[9]提出的一种使用多层注意力机制的远监督关系抽取模型,HMKN表示所提出的同时使用高速网络和多核卷积的远监督关系抽取方法;$\mathrm{HMKN_h}$为笔者提出模型的变形,仅使用了高速网络的包内包间注意力机制的远监督关系抽取方法;$\mathrm{HMKN_k}$为模型的另一种变形,仅使用了多核卷积的方法.

实验结果显示,HMKN模型的P@N值相较于之前的模型有了明显提高,在平均值有所提升的前提下P@100提升最为明显. 可以证明,通过多核卷积的方式可以对数据进行更深层次的提取,在分类器整体效果提高的同时对少部分示例更为敏感,分类效果更好.

表1 多种远监督关系抽取方法在NYT数据集上的实验结果

测试语句	P@N/%	BGRU + 3ATT	PCNN + ATT	PCNN + ATT + soft-label	$\mathrm{HMKN_h}$	$\mathrm{HMKN_k}$	HMKN
one	100	79.2	73.3	84.0	83.1	83.3	84.7
	200	73.1	69.2	75.5	73.9	76.5	76.2
	300	66.4	60.8	68.3	70.1	70.7	70.1
	平均	72.9	67.8	75.9	75.5	76.8	76.3
two	100	80.3	77.2	86.0	87.4	85.5	91.0
	200	75.9	71.6	77.0	76.5	80.5	79.8
	300	71.6	66.1	73.3	74.3	75.3	73.6
	平均	75.9	71.6	78.8	79.0	80.0	80.5
all	100	82.6	76.2	87.0	91.0	89.8	93.3
	200	77.7	73.1	84.5	81.5	83.4	83.7
	300	72.5	67.4	77.0	77.5	78.3	78.6
	平均	77.6	72.2	82.8	82.8	83.5	84.3

图 3 所示为几种关系抽取模型的 PR 曲线. 其中 Mintz[2]、MultiR[10] 和 MIMLRE[11] 是传统的基于特征的方法. PCNN + ATT 使用切分卷积神经网络和句子级别的注意力机制相结合的方法；PCNN + ATT + soft-label，是 Lin 等[6] 对 PCNN + ATT 的改进方法. BGRU + 3ATT 是李等[9] 提出的一种使用多层注意力机制的远监督关系抽取模型. 为了证明高速网络与多核卷积这 2 种方法在远监督关系抽取任务中的效果，通过对比实验的方法分别研究了单独使用其中一种方法对系统性能的影响.

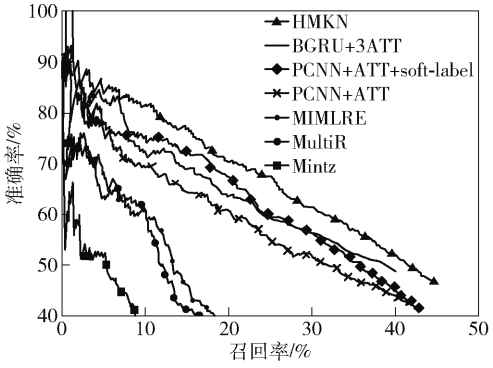

图 3　不同模型下的准确率和召回率

图 4 和表 2 显示了笔者提出的 $HMKN_h$、$HMKN_k$ 和 HMKN 这 3 种远监督关系抽取方法在 NYT 数据集上的 PR 曲线和 AUC 值. 结合图 4 和表 1、表 2 的结果可以证明，同时将高速网络和多非对称卷积方法加入神经网络中的效果比单独使用其中一种技术的效果要好. 通过 $HMKN_h$ 与 HMKN 实验结果的对比，说明多非对称核卷积的方法能够提高远监督任务的特征提取效果，并有助于提高分类器的性能. HMKN 相较于 $HMKN_k$ 的实验性能得到了提高，证明了通过高速网络的思想允许部分信息直接通过神经网络各层的做法可使特征和梯度的传递更加有效，同时减少了参数的数量.

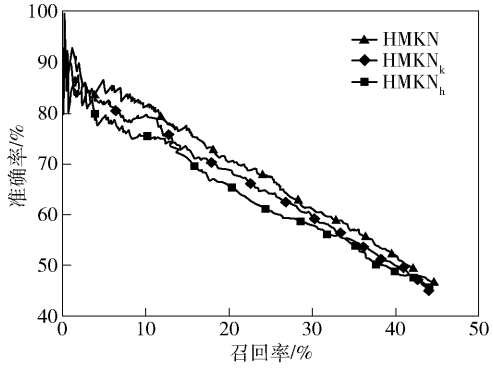

图 4　HMKN 模型和模型变体的准确率和召回率

表 2　HMKN 模型及模型变体的 AUC 值对比

模型	AUC
$HMKN_h$	0.403
$HMKN_k$	0.413
HMKN	0.419

3.4　模型复杂度分析

HMKN 通过多个非对称的卷积核实现更充分的跨句子特征提取. 模型的卷积操作的时间复杂度定义为 $O\left(\sum_{l=1}^{D} M_l^2 K_l^2 C_{l-1} C_l\right)$. 其中：$D$ 为神经网络的卷积层数；C_l 为第 l 个卷积层的卷积核个数；M 为每个卷积核输出特征图的边长；K 为每个卷积核的边长. 卷积操作的复杂度主要由模型的深度决定. HMKN 模型通过将 3 个卷积操作的结果相加，并没有增加每个卷积操作的层数，对更大的卷积核输出的特征图尺寸会更小，所以多核操作没有明显增加模型的复杂度.

4　结束语

提出了一种基于高速多核网络的远监督关系抽取模型，使用高速网络和多核卷积方法深度提取文本语义，并基于高速网络的包内注意力机制和包间注意力机制进行关系抽取，有效解决了远监督关系抽取任务中普遍存在的包噪声和句子噪声问题，提高了关系抽取的效率.

参考文献：

[1] Riedel S, Yao Limin, McCallum A. Modeling relations and their mentions without labeled text[J]. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2010(part 1): 148-163.

[2] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//International Joint Conference on Association for Computational Linguistics. Singapore: ACL and AFNLP, 2009: 1003-1011.

[3] Ding Xiaohan, Guo Yuchen, Ding Guiguang, et al. ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks[J]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019(3): 1911-1920.

[4] 冀振燕，宋晓军，皮怀雨，等. 基于深度学习的融合多源异构数据的推荐模型[J]. 北京邮电大学学报，2019，42(6)：35-42.

Ji Zhenyan, Song Xiaojun, Pi Huaiyu, et al. Recommended model for fusing multi-source heterogeneous data based on deep learning[J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(6): 35-42.

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

[6] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 2124-2133.

[7] Ye Zhixiu, Ling Zhenhua. Distant supervision relation extraction with intra-bag and inter-bag attentions[J]. Proceedings of NAACL-HLT, 2019(6): 2810-2819.

[8] Liu Tianyu, Wang Kexiang, Chang Baobao, et al. A soft-label method for noise-tolerant distantly supervised relation extraction[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017: 1790-1795.

[9] 李浩, 刘永坚, 解庆, 等. 基于多层次注意力机制的远程监督关系抽取模型[J]. 计算机科学, 2019, 46(10): 252-257.
Li Hao, Liu Yongjian, Xie Qing, et al. Distant supervision relation extraction model based on multi-level attention mechanism[J]. Computer Science, 2019, 46(10): 252-257.

[10] Raphael Hoffmann, Zhang Congle, Ling Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland: ACL, 2011: 541-550.

[11] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: ACL, 2012: 455-465.