文章编号:1007-5321(2018)01-0065-05 **DOI**:10.13190/j.jbupt.2017-127

融合卷积神经网络和重启随机游走的实体链接方法

谭咏梅[1], 李晓光[1], 吕学强[2]

(1. 北京邮电大学 智能科学与技术中心, 北京 100876;
2. 北京信息科技大学 网络文化与数字传播北京市重点实验室, 北京 100101)

摘要: 提出了一种融合卷积神经网络和重启随机游走的实体链接方法. 该方法首先对文本中的指称进行识别,然后生成指称的候选实体集,随后使用融合卷积神经网络和重启随机游走的实体链接方法对候选实体进行选择,最后对在知识库中无对应实体的指称进行聚类. 该方法在 TAC-KBP2016 的实体识别与链接评测数据集上的 F_{CEAFm} 值为 0.652,2016 年评测第 1 名的 F_{CEAFm} 为 0.643,实验结果表明,使用融合卷积神经网络和重启随机游走的实体链接方法能够有效地进行实体链接.

关 键 词: 实体链接; 卷积神经网络; 重启随机游走

中图分类号: TN911.22 **文献标志码**: A

An Entity Discover and Linking Approach Based on Convolutional Neural Network and Random Walk with Restart

TAN Yong-mei[1], LI Xiao-guang[1], LÜ Xue-qiang[2]

(1. Intelligence Science and Technology Center, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China)

Abstract: An entity linking approach based on convolutional neural network and random walk with restart was presented. This method first discovers the mentions in the text, after generates the mention candidate entity set, then selects the candidate entity using the entity linking approach based on convolutional neural network and random walk with restart and clusters the mentions those do not have the corresponding entity in the knowledge base. Our method F_{CEAFm} is 0.652 on the TAC-KBP2016 entity discovery and linking evaluation data set, and the first team is 0.643. The results show our method can effectively solve this problem.

Key words: entity linking; convolutional neural network; random walk with restart

针对实体链接,TAC(text analysis conference)连续多年设置了实体识别和链接(EDL, entity discovery and linking)的评测任务:给定一组文本,从文本中发现实体指称,并将这些发现的实体指称链接到 Freebase 知识库中对应的实体,且对在知识库中无对应实体的实体指称进行聚类[1].

目前实体链接主要面临 2 个问题:① 单一地使用实体指称和实体的文本信息或指称间相互关联的信息,两者没有同时得到有效使用;② 使用文本信息时严重依赖人工选取的特征,而现有的文本特征

收稿日期: 2017-07-03
基金项目: 国家自然科学基金面上项目(61671070); 网络文化与数字传播北京市重点实验室开放课题项目(ICDD201703)
作者简介: 谭咏梅(1975—), 女, 副教授, E-mail: ymtan@bupt.edu.cn.

往往不能有效地表达文本信息. 因此,笔者提出了一种融合卷积神经网络和重启随机游走的实体识别与链接方法,使用重启随机游走获取的知识库信息融合卷积神经网络获取的文本信息进行实体链接.

1 相关工作

研究者提出的解决实体链接的方法主要分为单一式实体链接和协同式实体链接2种方法[2]. 单一式的实体链接通过比较指称的上下文和候选实体在知识库中描述文本的相似度大小进行链接. He等[3]使用深度神经网络对指称的上下文和候选实体的描述文本信息进行编码,取代了文本的词袋向量表示形式,大大减少了向量的维度,同时避免了人工特征选择问题. Sun等[4]通过卷积神经网络使用指称的上下文和候选实体在知识库中的信息进行实体链接. Matthew等[5]使用卷积神经网络计算指称和实体的语义相似度进行链接. 因此,单一式的实体链接往往仅考虑指称的上下文信息和实体的描述文本信息,忽略了指称间的相互关联.

协同式的实体链接对同一文本中的全部实体指称一起进行实体链接,综合考虑多个指称间的语义关联,建立全局语义约束,继而更好地对指称进行链接. 李茂林[2]通过构造指称-实体图,并使用随机游走算法在图上得到平稳分布,选择权重最高的为目标链接实体. Guo等[6]通过重启随机游走算法在实体图上获取实体和文本统一的语义表达进行实体链接. Tan等[7]在Guo等基础上使用迭代的方式进行实体链接,每次会使用当前已经完成链接的指称对实体图进行裁剪,即将上次链接的信息传递到下一次计算中. 但如果当前指称链接到错误的实体,则错误的信息会向后传递,影响后续实体的链接. 上述协同式的实体链接方法在进行实体链接时,使用了文本指称间的相互关联信息,缺点是严重依赖文本中指称的识别步骤. 如果存在错误识别的指称,则错误的指称可能影响其他指称的链接结果.

2 实体链接

融合卷积神经网络和重启随机游走的实体链接方法首先对文本和知识库进行预处理,然后通过指称识别部分识别文本中的指称,随后候选实体生成部分获取指称的候选实体集,最后使用实体选择和指称聚类部分对指称进行链接和聚类,得到最终结果,如图1所示.

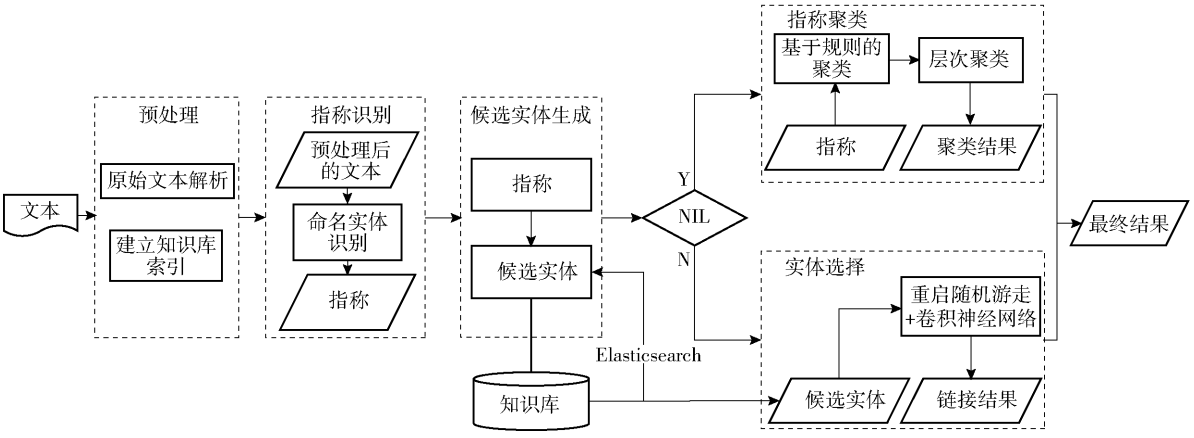

图1 融合卷积神经网络和重启随机游走的实体识别与链接框架

2.1 预处理

知识库是由互联网中无(半)结构的知识组织成的结构化知识集合,它描述的对象是实体以及实体间的关系,实体之间的关系用三元组(实体1,关系,实体2)来表示. 笔者使用的知识库为Freebase,其包含了约19亿条记录,如果使用传统的线性遍历查找,则时间开销会非常大. 笔者使用Elasticsearch①对知识库构建索引,Elasticsearch是一个实时分布式搜索和分析引擎,因此可以将每次搜索实体的时间降低到秒级,从而提高搜索候选实体的效率. 由于使用的评测数据集中存在一些无用和冗余的信息,所以对原始文本进行了解析,以提取有效的文本.

2.2 指称识别

使用Stanford NER②工具对文本中的实体指称

① https://www.elastic.co/

② http://nlp.stanford.edu/ner/

进行识别. Stanford NER 工具可识别文本中的大部分指称,但存在对缩写、简写、别名、昵称等形式的指称难以准确识别以及对普通地点名、便利设施识别效果较差等问题,因此,可通过外部资源构建的实体多样性词表来解决指称的缩写和简写等问题,然后通过词向量的相似性构造普通地点名和便利设施表来提高这 2 种实体的识别准确率.

2.3　候选实体生成

在生成候选实体时,一个指称存在过多的候选实体可能会存在数据稀疏的问题,影响实体链接的效果. 因此,候选实体生成应使候选实体集在包含正确链接实体的情况下尽可能地小. 笔者使用 Elasticsearch 搜索引擎的匹配算法以及指称类型与知识库中实体类型的对应关系生成并筛选指称的候选实体集.

2.4　实体选择

首先使用卷积神经网络对指称的上下文以及实体在知识库中的描述文本进行特征提取,然后拼接使用重启随机游走得到的指称语义特征 $\boldsymbol{F}(m)$ 和实体语义特征 $\boldsymbol{F}(e)$,再经过一个全连接层,最后计算距离得到链接结果. 实体选择模型如图 2 所示.

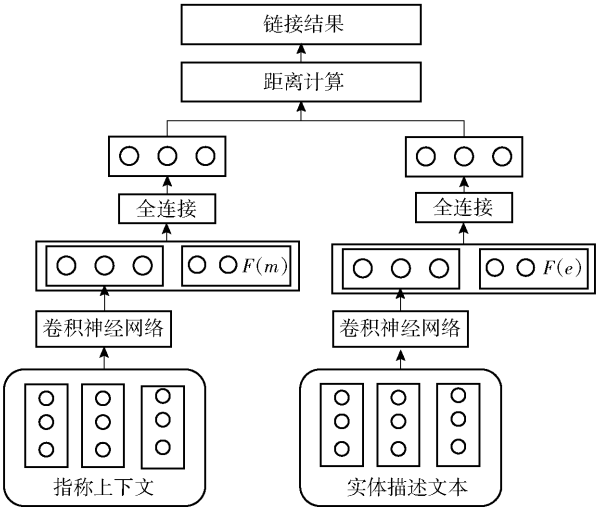

图 2　实体选择模型

2.4.1　语义特征

语义特征参照了 Guo 等[6]和 Tan 等[7]的工作.

1) 指称-实体图的构建

指称—实体图的构造方式为 $G=(V,E)$. 其中:V 为指称、候选实体以及扩充实体的集合,扩充实体为与 2 个候选实体都有关的实体;E 为候选实体间的边以及指称和候选实体间的边.

重启随机游走中的转移矩阵用 $\boldsymbol{T}$ 表示,$\boldsymbol{T}$ 中的元素 t_{ij} 表示从节点 e_i 访问节点 e_j 的概率. t_{ij} 的计算公式[7]为

$$t_{ij}=\frac{w_{ij}}{\sum_{k\in N(\boldsymbol{e}_i)} w_{ik}} \tag{1}$$

其中:$N(e_i)$ 为与节点 e_i 直接相连的节点集合,w_{ij} 为节点 e_i 与节点 e_j 之间的权重.

2) 实体的语义特征

计算一个目标实体 e_i 的语义特征时[7],需要重启随机游走过程始终从 e_i 重启. 因此通过设置初始向量 s 的分量 $s_i=1$,其他分量 $s_j(j\neq i)=0$ 可满足这一条件. 得到初始向量后,使用重启的随机游走得到实体 e 的语义特征向量 $\boldsymbol{F}(e)$:

$$\boldsymbol{F}(e)=\alpha(I-c\boldsymbol{T})^{-1}\boldsymbol{s},\quad c=1-\alpha \tag{2}$$

其中 α 为重启概率.

3) 指称的语义特征

计算指称的语义特征时[7],将该指称的候选实集 Q_m 设置为初始节点,并计算候选实体 e_i 在初始向量 $\boldsymbol{s}$ 中的对应权重 $s_i=P(m\rightarrow e_i)$,计算公式为

$$P(m\rightarrow e_i)=\frac{R(m,e_i)}{\sum_{e_k\in Q_m} R(m,e_k)} \tag{3}$$

其中 $R(m,\ e_i)$ 表示指称 m 和实体 e_i 之间的相似度. 根据词袋模型将指称的上下文和实体的描述文本分别转换为向量 $\boldsymbol{m}$ 和 $\boldsymbol{e}$,则 $R(\boldsymbol{m},\ \boldsymbol{e})$ 计算方式[7]为

$$R(\boldsymbol{m},\boldsymbol{e})=\frac{\boldsymbol{me}}{|\boldsymbol{m}||\boldsymbol{e}|} \tag{4}$$

得到初始向量后,使用重启随机游走得到指称 m 的语义特征向量 $\boldsymbol{F}(m)$:

$$\boldsymbol{F}(m)=\alpha(I-cT)^{-1}\boldsymbol{s},c=1-\alpha \tag{5}$$

2.4.2　卷积神经网络结构

获取指称的上下文和实体描述文本特征的卷积神经网络结构如图 3 所示.

首先通过词向量层将文本映射为数字矩阵,然后依次经过卷积层 1、池化层 1、卷积层 2 和池化层 2,接着通过展平操作将特征图转换为一维向量,最后经过 dropout 层和全连接层输出文本的特征.

2.4.3　模型训练

损失函数使用了孪生网络的对比损失函数来训练网络模型参数[8].

$$d(m,e)=\|v(m)-v(e)\|_2 \tag{6}$$

$$L=yd^2+(1-y)\max(1-d,0)^2 \tag{7}$$

其中:$v(e)$ 和 $v(m)$ 为卷积神经网络输出拼接语义特征后经过全连接层输出的结果;如果实体 e 为指称

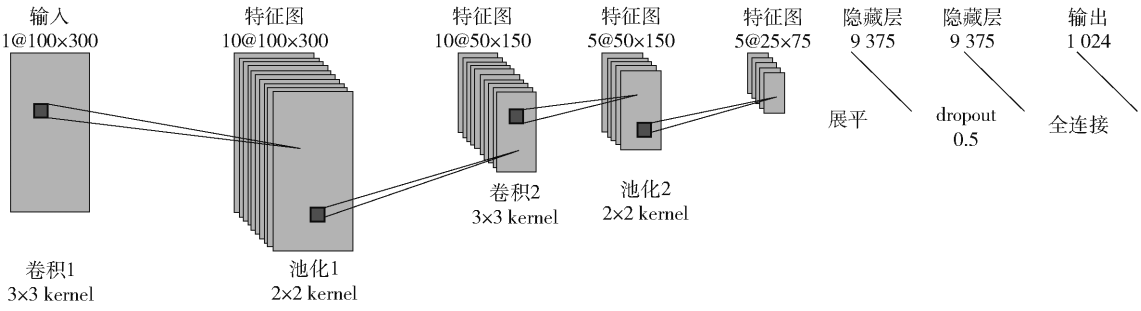

图3 卷积神经网络网络结构

m 的链接目标,则 $y=1$,相反 $y=0$;L 为损失值;d 为指称和候选实体的距离.

2.5 指称聚类

首先参照李茂林[2]使用人工编写的规则对指称进行粗划分,如果2个指称 m_i 和 m_j 的命名实体类型相同,且满足以下一个条件,则将其暂时划为一簇:①指称 m_i 和 m_j 有相同的表面字符串;②m_i 和 m_j 的前一部分匹配,或者是 m_j 和 m_i 的前一部分匹配;③m_i 和 m_j 的后一部分匹配,或者是 m_j 和 m_i 的后一部分匹配. 若指称 m_i 和 m_j 的命名实体类型不同,则不能聚为一类.

然后通过将实体指称对应的上下文转换为向量表示,使用层次聚类算法进行指称聚类.

3 实验

3.1 实验数据

采用2016年TAC-KBP的EDL评测任务的实验数据①,使用的知识库为Freebase(2015年1月份版本).

3.2 评价指标

使用TAC-KBP中EDL评测任务的CEAFm(mention ceaf)[1]指标对链接和聚类结果进行评价. 对于链接部分,该指标要求指称在文本中的位置、指称类型以及指称对应的知识库中的实体id和官方所给的标准答案完全一致,否则该指称链接不正确. 对于聚类部分,该指标要求表示相同实体的聚为一类,即具有相同的id.

3.3 实验结果及分析

所提方法与2016年参加TAC-KPB评测取得前3名的参赛队伍[1]进行了对比,结果如表1所示.

由表1可以得出,所提出的融合卷积神经网络和重启随机游走的实体识别与链接方法在2016年TAC-KBP的EDL评测语料优于其他方法. Tan等[7]提出的基于重启随机游走的实体链接方法通过构建与文本有关的实体图来进行实体链接,该方法仅仅考虑文本中识别出的指称相关信息,对指称识别有较高的要求,且使用迭代的方法进行链接,存在错误累积的问题. 笔者充分使用指称的上下文和知识库中实体的信息进行建模,提升了实体链接的性能.

表1 EDL的实验结果

实验方法	P_{CEAFm}	R_{CEAFm}	F_{CEAPm}
所提方法	**0.772**	**0.564**	**0.652**
评测第1名	0.734	0.572	0.643
评测第2名[6]	0.757	0.553	0.639
评测第3名	0.798	0.531	0.638

不同模型下的实验结果如表2所示.

表2 模型结果对比

实验方法	P_{CEAFm}	R_{CEAFm}	F_{CEAFm}
卷积神经网络	0.749	0.522	0.615
重启随机游走	0.757	0.553	0.639
所提方法	0.772	0.564	0.652

由表2可以得出,卷积神经网络的效果在3个模型中最差. 使用指称的局部信息进行链接,需要丰富的指称局部信息,如"联合国秘书长潘基文强烈谴责对特派团营地的袭击",其中局部信息"联合国秘书长"对指称"潘基文"的链接结果贡献了非常重要的信息. 重启随机游走模型使用全局信息进行协同式链接,提高了链接效果. 融合卷积神经网络和重启随机游走的模型在全局信息的基础上引入局部重点关注信息,效果在3种模型中最优.

4 结束语

针对实体链接问题,提出了一种融合卷积神经网络和重启随机游走的实体识别与链接方法. 该方

① http://nlp.cs.rpi.edu/kbp/2016/index.html

法在 TAC-KBP2016 的实体链接数据集上的 F_{CEAFm} 值为 0. 652,2016 年 KBP 评测参赛第 1 名的 F_{CEAFm} 为 0. 643,实验结果表明了所提方法的有效性.

所提方法尚需改进的方面有:①使用了 Stanford NER 识别工具对文中的指称进行识别,但是此工具对地名、设施名识别效果并不理想,虽然增加了词表来解决此问题,但同时也增加了人工成本. 目前长短记忆神经网络模型在序列标注问题上的优势较为明显,后期可使用此模型解决实体识别的问题;②知识库中一些实体不存在描述文本,该方法可能退化为使用重启随机游走解决实体链接问题.

参考文献:

[1] Ji Heng, Nothman Joel, Dang Hoa Trang. Overview of TAC-KBP2016 tri-lingual EDL and its impact on end-to-end cold-start KBP[C]// TAC. Gaithersburg, Maryland: National Institute of Standards and Technology, 2016: 21-35.

[2] 李茂林. 基于主题敏感的重启随机游走实体链接方法[J]. 北京大学学报(自然科学版), 2016, 52(1): 17-24.
Li Maolin. An entity linking approach based on topic-sensitive random walk with restart[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 17-24.

[3] He Zhengyan, Liu Shujie, Song Yang, et al. Efficient collective entity linking with stacking[C]// EMNLP. Seattle: Association for Computational Linguistics, 2013: 426-435.

[4] Sun Yaming, Lin Lei, Tang Duyu, et al. Modeling mention, context and entity with neural networks for entity disambiguation[C]// International Conference on Artificial Intelligence. Buenos Aires: AAAI Press, 2015: 1333-1339.

[5] Matthew Francis, Durrett Greg, Klein Dan. Capturing semantic similarity for entity linking with convolutional neural networks[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016: 1256-1261.

[6] Guo Zhaochen, Barbosa Denilson. Robust entity linking via random walks[C]// ACM International Conference on Information and Knowledge Management. Shanghai: ACM, 2014: 499-508.

[7] Tan Yongmei, Li Xiaoguang, Zheng Di. BUPTTeam participation at TAC 2016 knowledge base population[C]// TAC. Gaithersburg, Maryland: National Institute of Standards and Technology, 2016: 55-60.

[8] Hadsell Raia, Chopra Sumit, Lecun Yann. Dimensionality reduction by learning an invariant mapping[C]//IEEE Computer Society Conferenc on Computer Vision and Pattern Recognition. New York: IEEE Computer Society, 2006: 1735-1740.