doi:10.3772/j.issn.2095-915x.2018.03.006

科技情报对象关系抽取的技术选择

刘秀磊[1] 王延飞[2] 刘思含[1] 李红臣[3]

1. 北京信息科技大学数据科学与情报分析实验室 北京 100101;
2. 北京大学信息管理系 北京 100871;
3. 国家安全生产监督管理总局通信信息中心 北京 100013

摘 要 科技情报数据与日俱增，呈现海量、多源、异构的特性。针对上述特点，知识图谱能较深入地分析科技情报，实现对科技情报对象的感知和刻画。科技情报对象关系的抽取是知识图谱构建的关键步骤。本文在总结分析现有关系抽取技术和传统神经网络模型的基础上，提出一种基于长短期记忆神经网络和卷积神经网络深度机器学习的关系抽取技术，并通过实验证明其具有较好的性能，与其他运用神经网络的关系抽取技术相比，该技术的准确率、召回率和*F*值均有提高，运用该技术有利于增强对科技情报对象关系的感知刻画能力。

关键词： 关系抽取；科技情报；长短期记忆神经网络；卷积神经网络
中图分类号： G35
开放科学（资源服务）标识码（OSID）

Technology Choice of Relation Extraction of Scientific and Technical Objects

LIU Xiulei[1] WANG Yanfei[2] LIU Sihan[1] LI Hongchen[3]

1. Laboratory of Data Science and Information Analysis, Beijing Information Science and Technology University, Beijing 100101, China;
2. Dept. of Information Management, Peking University, Beijing 100871, China;
3. State Administration of Work Safety Communication Information Center, Beijing 100013, China

基金项目： 国家重点研发计划课题“基于云技术的煤矿典型动力灾害区域监控预警系统平台”（2016YFC0801407）；国家自然科学基金“物联网搜索中异构本体的语义融合研究”（61601039）；北京信息科技大学学校科研基金项目“本体匹配中词法分析关键技术研究”（1625008）；网络文化与数字传播北京市重点实验室开放课题“基于词法分析和语义分析的本体匹配研究”(ICDDXN006)；北京信息科技大学软件工程专业学位点建设项目（5121723402）。
作者简介： 刘秀磊（1981–），博士，副教授，研究方向：语义web、知识图谱、大数据情报分析，E-mail: xiuleiliu@hotmail.com；王延飞（1965–），通讯作者，博士，教授，研究方向：情报研究；刘思含（1993–），硕士研究生，研究方向：自然语言处理、知识图谱。

doi:10.3772/j.issn.2095-915x.2018.03.006

Abstract The growing scientific and technical intelligence object data presented the characteristics of massive, multi-sourced and heterogeneous. For the above characteristics, knowledge graph can deeply analyze the scientific and technical intelligence and realize the perception, portrayal and merging of scientific and technical intelligence objects. Relation extraction of scientific and technical intelligence object is the key step in the construction of knowledge map. This paper proposes a relation extraction technology based on the long short-term memory and convolutional neural network deep machine learning. The results of experiment indicate that the new method obtain better performances compared with other neural networks extracted from relation including the accuracy, recall rate and F value. This new technology can help to enhance the portrayal and integration of science and technology intelligence objects.

Keywords: Relation extraction; scientific and technical intelligence; long short-term memory; convolutional neural network

1 引言

随着科技情报数据逐渐呈现海量、多源、异构的特性，传统的情报分析方法在明确学科群发展规律、展示情报对象结构关系、探索新兴领域方向等方面面临严峻压力，因此，情报学者尝试将知识图谱用于科技情报分析，在数据融合基础上实现对科技情报对象关系的感知和刻画，以期有效解决分析任务压力问题。

王伟军等人从课题研究领域、技术支持、专家建议预见和相关成果扩散等方面详细阐述和论证了知识图谱在情报分析应用的可行性和必要性[1]。李慧贞借助知识图谱从微观、中观、宏观三个角度分析了学科结构和演化[2]。田恬提出一种基于CiteSpace的知识图谱计量分析方法，深入分析了知识的结构和特征[3]。王海燕等人利用知识图谱将科技文献转变为可视化图像，挖掘出某一科学领域的知识演变过程[4]。焦晓静等人提出了知识图谱在科技情报领域应用的具体节点[5]。王姣和孙林使用知识图谱的技术总结分析了国内2005–2015年间情报分析领域的研究现状，指出国内情报分析呈现出研究多领域合作、系统性分析、处理数据化、网络化监测、情报人员培训五大特点[6]。李雅等人使用聚类分析和高频关键词分类排序等知识图谱方法分析了乙酰甲喹研究生长点、趋势及其实验检测指标[7]。徐珂珂使用“专利-特征项”矩阵构建专利文献关键技术的知识图谱框架[8]。张同同基于知识图谱和关键词共现分析并揭示了科技情报研究机构的科研发展现状[9]。

上述研究阐述并实现了知识图谱在科技情报领域应用的可行性，但是均未对知识图谱构建的关键步骤中关系抽取技术进行详细地说明和研究。本文尝试总结并分析现有的关系抽取技术，在这些研究基础上提出一种基于长短期记忆神经网络和卷积神经网络深度机器学习的

关系抽取技术，使用前向学习和反向学习的方法并将文本信息用最短依存路径表示，以求更好地理解记忆输入的内容。同时，为了提高训练模型的可靠性，借助依存分析提取出文本主要的语法结构。最后，通过实验证明本文提出的关系抽取模型具有更好的识别能力和扩展性能，并对未来科技情报对象关系抽取技术研究提出展望。

2 国内外关系抽取研究现状

关系抽取的任务分为分类关系抽取和非分类关系抽取。分类关系是指对象之间的层次关系，可解释为一种“is-a”关系。Ann等人使用CBC（clustering by Committee）的聚类方法将抽取到的对象组织成层次结构[10]。非分类关系是指除了分类关系之外的所有关系，是当前关系抽取的研究重点。句法分析、依存分析等深层自然语言处理技术较多应用在非分类关系的抽取中。文献[11]和[12]采用关系三元组的形式，通过识别出核心动词挖掘对象间关系。Quan Fang等人根据对象间的共现信息提出一种基于规则映射的关系抽取方法[13]。文献[14]和[15]中，学者综合共现信息、字符串匹配信息、上下文信息通过计算对象间的某种相似度值并根据计算结果对其分配对应类型的关系。一些研究针对特定关系进行抽取，比如文献[16]和[17]中，学者首先构造了一些基于特定关系的意图查询，然后将其提交给搜索引擎。搜索引擎查询网页中具有对应关系的语料资源。查找到的语料会根据网页中的html标记和意图查询的格式进行过滤，并从中抽取候选的关系。最后基于特定关系在自然语言表述中的特点和汉语的构词规律，选出目标关系。Fanghuai Hu等人提前对要抽取的关系进行预定义，利用词表相似度、词性、共现信息、tf-idf等特征，通过SVM、最大熵等分类器，为新的分类目标分配对应的关系类型[18]。其本质是将关系的抽取任务看作分类任务。

3 基于 LSTM 和 CNN 深度机器学习的科技情报对象关系抽取技术

上述关系抽取的方法主要基于自然语言处理和机器学习技术，所以存在标注数据有限，人工标注劳动力大，特征构建能力有限，隐含关系挖掘不够等缺陷。随着深度学习的发展并在图像处理领域已取得良好效果，相关研究开始在自然语言处理领域进行探索。学者已在关系分类的研究上取得了很好的成果。例如，文献[19]和文献[20]分别使用递归神经网(recursive neural network，简称RNN)和卷积神经网络（convolutional neural network，简称CNN）进行关系分类的研究。传统的神经网络存在一些缺陷，如需要通过大量的非线性变换，参数过多，容易出现过拟合现象；层数不够，提取特征信息不够全面；梯度消失等问题。针对传统关系抽取方法以及传统神经网络模型的缺陷，本文提出一种基于长短期记忆神经网络（long short-term memory，简称LSTM）和卷积神经网络的模型的关系抽取方法。

3.1 整体网络结构

整体的网络结构如图1所示。

doi:10.3772/j.issn.2095-915x.2018.03.006

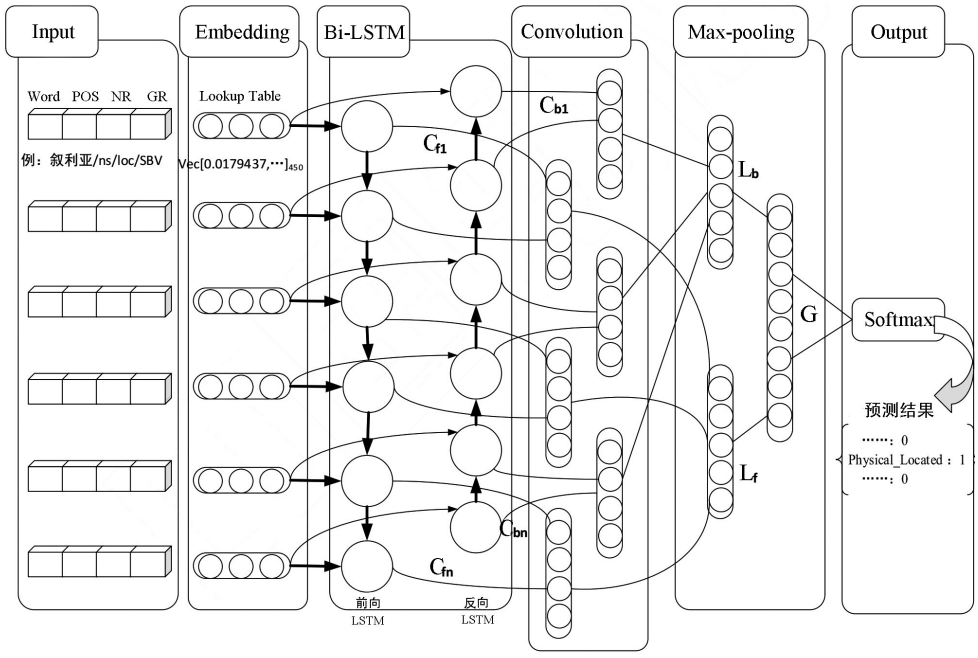

图 1 网络结构

第一层为输入层，用于对文本进行依存分析、分词等预处理，得到所需的词序列及词汇表示特征；第二层为向量表示层，使用word2vec训练的向量模型表示词汇；第三层为循环神经网络，将处理好的语料输入LSTM单元进行训练；第四层为卷积神经网络，把LSTM 单元的记忆输出按正向和反向的输出结果通过卷积神经网络得到两个不同输入方向的表示；之后对所得的表示进行向量级联；第五层为池化层，使用最大池操作得到这条输入语料的最终向量表示；第六层为输出层，使用集成softmax函数计算出语料的预测类别。

3.2 数据预处理和词向量表示

数据预处理工具可以使用LTP-Cloud，对文本进行分词、词性标注、依存分析等，并获得最短依存路径、对象类别和语法关系。然后，生成形如$[w_1,w_2,\ldots\ldots,w_n]$词序列，其中$w$的结构表示为$[Word,POS,NR,GR]$。词向量$Word$使用word2vec生成映射词表得到其向量表示。词性标记POS是对词汇类别的粗粒度表示，如名词noun，动词verb等。命名实体识别(Named Entity Recognition, 简称NER)是对词序列中的人名、地名、机构名等实体进行定位和标记。NR用于概括词汇实体的概念表述，对于非命名实体，则采用零填充方法补齐维度进行表示。语法关系CR是通过使用依存句法分析识别句子中的语法成分，并据此分析出句法结构。

3.3 双通道循环神经网络

传统的循环神经网络（recurrent neural networks，简称RNN）存在梯度消失、爆炸、历史信息损失等问题。因此，Hochreiter等人提出LSTM单元用于解决上述问题[21]。LSTM实质是一个自适应的门机制，能够决定记忆单元保留上一级记忆状态和提取当前输入特征的程

度。LSTM由以下四部分构成：输入门i，遗忘门f，输出门o和记忆单元c，其结构如图2所示。

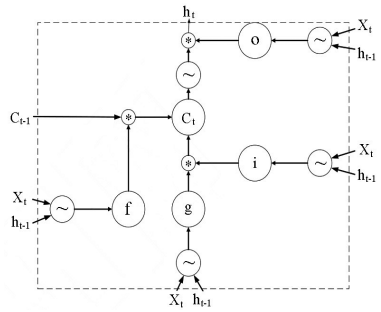

图 2 LSTM 单元

对于三个门的计算，下标t表示时间状态t时的输入，其中依赖上一步隐层状态向量h_t，当前输入x_t。

$$i_t=\sigma(W_i\cdot x_t+U_i\cdot h_{t-1}+b_i) \quad （公式1）$$

$$f_t=\sigma(W_f\cdot x_t+U_f\cdot h_{t-1}+b_f) \quad （公式2）$$

$$o_t=\sigma(W_o\cdot x_t+U_o\cdot h_{t-1}+b_o) \quad （公式3）$$

其中，特征向量的提取用g_t计算。

$$g_t=\tanh(W_g\cdot x_t+U_g\cdot h_{t-1}+b_g) \quad （公式4）$$

当前记忆单元c_t通过计算之前的单元信息c_{t-1}和候选内容g_t得到，并且受输入门i_t和遗忘门g_t影响。

$$c_t=\left[i_t\otimes g_t+f_t\otimes c_{t-1}\right] \quad （公式5）$$

此时，LSTM的输出值即为循环神经网络的隐层状态h_t，具体表示为：

$$h_t=o_t\otimes\tanh\left(c_t\right) \quad （公式6）$$

最短依存路径是由一些形如$w_a\xrightarrow{r_{ab}}w_b$的依存单元子树结构组成。路径之间的词汇信息位于关系发生时关键词根节点到左右两边实体节点之间，包含文本表达的主要信息以及词汇间的隐含信息，同时删除冗余的噪声信息。

在隐藏层中，本文使用双通道循环神经网络，通过对正向词序列和反向词序列的训练，更好的提取到文本的主要语义信息。每个通道的LSTM单元的输出为词汇w的信息表示，其中w为经word2vec训练后得到的形如$(x_0,x_1,\ldots\ldots,x_n)$的词向量，词序列的顺序按照最短依存路径的左右分支顺序。依存单元的输出为$d_b=\left[h_a\otimes h_b\right]$，其中$h_a$、$h_b$为$w$的隐层计算结果。为了能最大限度的挖掘出更多隐藏特征，使用保留每次记忆单元的结果d_{ab}作为当前时刻的局部特征。并将局部特征组合后的结果传递至卷积层进行训练。

3.4 卷积神经网络

CNN是由Yann LeCun和他的同事于1998年提出[22]。相较于一般的神经网络，CNN在对数据的操作上较为简单。不仅可以实现隐式的特征提取，显式特征提取准确率也较高。由于CNN的权值共享，减少了网络的训练参数，因此降低了神经网络的复杂度，训练速度更快。本文使用的CNN结构如图3所示。

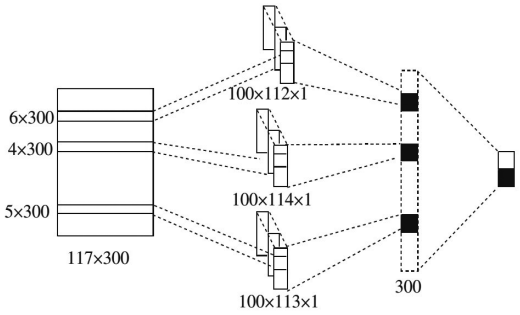

图 3 CNN 模型结构

记忆单元的结果d_{ab}作为卷积神经网络的输入，单个通道内的d_{ab}级联为该顺序文本特征信息。特征矩阵的宽为LSTM输出的特征宽度，长为文本顺序输入时的句子长度。通道的d_{ab}经过卷积神经网络的计算得到局部特征L_{ab}，单通道的局部特征组合L可以理解为单向特征表示。L的计算表示为：

doi:10.3772/j.issn.2095-915x.2018.03.006

$$L=\sum_{n=1}^{l}\left[L_{ab}\right]_n \quad \text{（公式7）}$$

其中，卷积计算L_{ab}，W_{con}是卷积层的权重矩阵，b_{con}是其隐层中的偏置量，非线性激励函数f选择tanh函数。

$$L_{ab}=f\left(W_{con}\cdot\left[h_a\otimes h_b\right]+b_{con}\right) \quad \text{（公式8）}$$

3.5 最大池化

虽然通过卷积神经网络得到的特征可以用于训练分类器，但是计算量巨大。为了解决这个问题，采用最大池化的方法。最大池化可以提高特征探测器的通用性，因为隐层的稀疏表示被划为若干不重叠的子区域，并去除这些子区域中的最大值。由于重构的过程中，隐式编码使神经元的平均数量减少，所以弱化了隐层单元或权重给予L的正则化条件。其过程如图4所示。

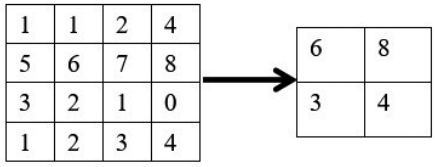

图 4 池化示意图

首先确定池化区域的大小；然后，把获取到的卷积特征划分成若干不相交的子区域；最后，取每个区域的最大特征作为池化后的卷积特征。本文的全局特征G为通道的依存单元计算结果的组合，即$G=\max_{d=1}^{D}L_d\ (10)$。若将池化窗口尺寸设置为p，特征映射分隔成n个窗口，则选取窗口内最佳匹配结果$g=\max(p_0,p_1,\ldots\ldots,p_{n-1}]$代表窗口的全局特征。

3.6 集成分类器

输出层采用集成softmax回归对关系进行预测分类，目的是使样本属于不同关系类型的概率和为1，该集成分类器是基于子空间的，可以处理一个样本有多个输入。给定训练数据$X=\{X^{(1)},X^{(2)},\ldots\ldots,X^{(K)}\}$和对应关系类别标签$\{l_1,l_2,\ldots\ldots,l_n\}$，其中$X^{(k)}=\left\{x_1^{(1)},x_2^{(2)},\ldots\ldots,x_k^{(k)}\right\}\in R^{mk}$是训练数据的第$k$个子空间。给定一个测试数据$x=\{x^{(1)},x^{(2)},\ldots\ldots,x^{(K)}\}$，其中$x^{(k)}$是$x$的第$k$个子空间。

集成softmax函数对样本x的分类结果如下[23]：

$$h_{\theta(x)}=\begin{bmatrix} p\left(l_i=1\mid x^{(1)};\theta^{(1)}\right)+\ldots\ldots+p\left(l_i=1\mid x^{(K)};\theta^{(K)}\right)\\ p\left(l_i=2\mid x^{(1)};\theta^{(1)}\right)+\ldots\ldots+p\left(l_i=2\mid x^{(K)};\theta^{(K)}\right)\\ \cdot\\ \cdot\\ \cdot\\ p\left(l_i=C\mid x^{(1)};\theta^{(1)}\right)+\ldots\ldots+p\left(l_i=C\mid x^{(K)};\theta^{(K)}\right)\end{bmatrix}$$

$$=\sum_{k=1}^{K}\frac{1}{\sum_{j=1}^{C}e^{\left(\theta_j^{(k)}\right)^{Tx^{(k)}}}}\begin{bmatrix} e^{\left(\theta_1^{(k)}\right)^{Tx^{(k)}}}\\ e^{\left(\theta_2^{(k)}\right)^{Tx^{(k)}}}\\ \cdot\\ \cdot\\ \cdot\\ e^{\left(\theta_C^{(k)}\right)^{Tx^{(k)}}}\end{bmatrix} \quad \text{（公式9）}$$

其中，$\theta_j^{(k)}$是第j类第k个特征子空间的参数。

4 实验结果分析

4.1 实验描述

本文所使用的语料为某专利公司计算机领域的专利文献共10000篇，文献30000篇，科研机构主页2000篇作为输入。预处理后，将9000条作为训练语料。关系类型的定义参照ACE05

的标准分为10种，具体关系标签如表1所示。

表 1　关系类别标签

Relation Labels
位于关系(Physical_Located)
相邻关系(Physical_Near)
社会关系(Person_Socical)
亲属关系(Person_Family)
层次关系(Part_Whole)
占有关系(Org-Aff_Ownership)
雇佣关系(Org-Aff_Employment)
成员关系(Org-Aff_Membership)
施事关系(Agent-Artifact)
其他关系(Others)

文本分别存储在不同类别的训练集中，通过独热编码添加训练标签。通过softmax函数计算出属于该文本的n维向量，n的取值为关系类别的数量。向量取值范围为0~1之间，即$p\in[0,1]$，p为此文本属于某类关系的概率，选取p最大值作为其关系类别。实验所用的词向量生成维度为300维，输入特征维度为{300,50,50,50}，卷积维度为100维，隐层单元数量为300，l_2惩罚权重系数为10^{-5}。为避免过拟合，使用dropout策略，比率为0.3，学习效率为0.002。实验经过600轮训练，进行10折交叉验证，采用正确率（P）、召回率（R）、F值作为性能评价指标。

4.2 实验分析

相较于传统CNN、Bi-LSTM模型，根据评价指标对比了每种关系的抽取效果和整体抽取效果分别如图5、6、7和表2所示。

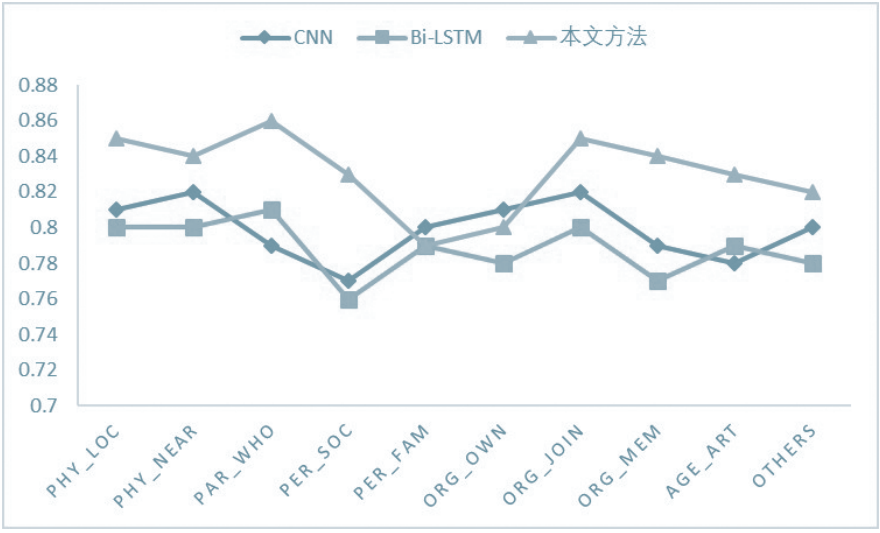

图 5　关系抽取准确率比较

从上述图表中，可以看出本文提出的抽取方法仅在某一或两种关系的抽取效果上表现欠佳，对于绝大多种关系抽取的准确率、召回率和F值均有提升。对比可以看出，本文提出的方法对于Part_Whole和Person_Socical这两种关系在准确率、召回率和F值上都明显高于其他种类关系。通过分析语料发现，这两种关系在实体标注和依存分析方面具有较鲜明特点，所以特征更加准确可靠。而且在关系出现的节点上，具有高频词汇。表2计算了关系抽取的整体平均效果，证明了本文提出的方法总体上达到了预期效果。

doi:10.3772/j.issn.2095-915x.2018.03.006

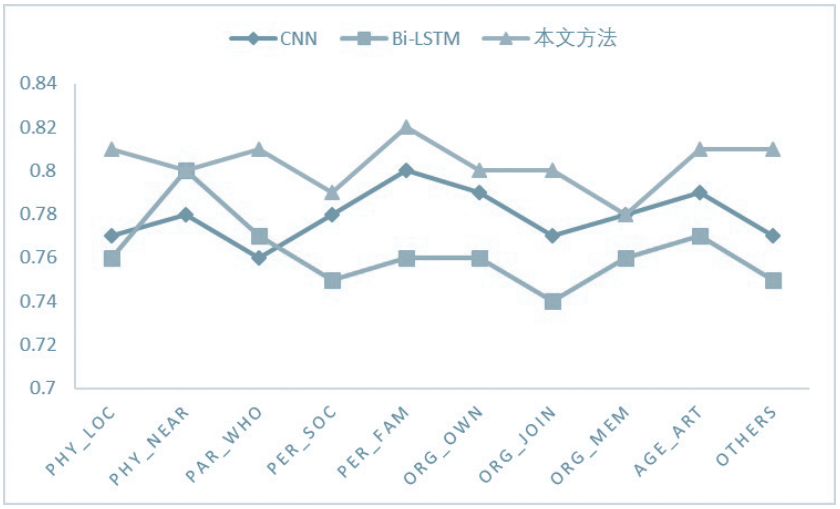

图 6 关系抽取召回率比较

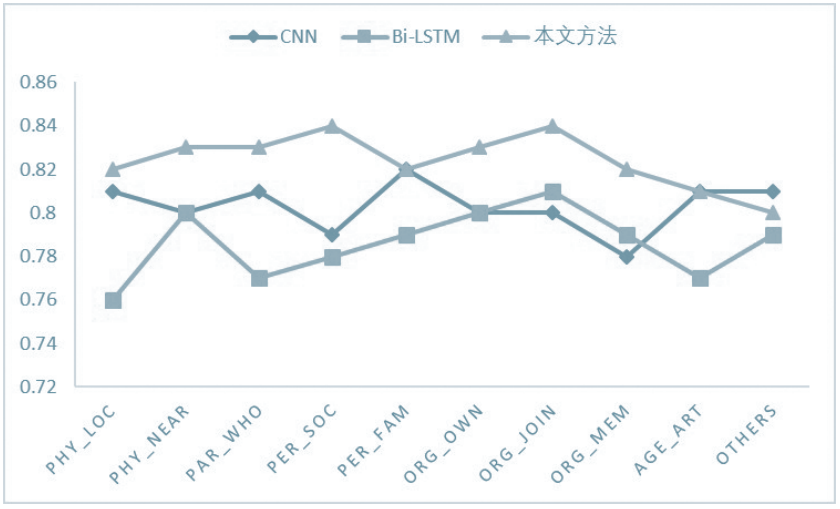

图 7 关系抽取 F 值比较

表 2 关系抽取实验结果对比

模型	特征集	准确率	召回率	*F*值
CNN	[*SPT*+*pos*+*ner*+*gr*]	81.3	77.9	80.2
Bi-LSTM	[*SPT*+*pos*+*ner*+*gr*]	80.5	76.5	79.0
本文方法	[*SPT*+*pos*+*ner*+*gr*]	85.4	81.1	83.1

5 总结与展望

科技情报数据海量、多源、异构的特点导致情报分析的方式从信息获取存储管理、分类、索引、检索、聚类、人机交互技术这种传统方式向知识抽取、知识图谱等情报分析过程转变。相关学者为了实现对于科技情报对象的感知、刻画和融合，开始将知识图谱技术融入科技情报分析领域，而科技情报对象关系抽取作为知识图谱构建的关键步骤，将被越来越多的学者关注。本文针对目前知识图谱与科技情报领域相结合的研究现状以及现有的关系抽取

方法的缺陷，提出了一种基于LSTM和CNN的科技情报对象关系抽取技术，并通过实验证明了该方法的性能优越性。但是，还存在一些不足之处。本文仅对预定义的关系进行抽取，并未对开放式关系进行研究。目前，在开放式关系抽取研究上，还存在消歧问题。开放式关系的抽取主要基于核心动词的提取，由于中文文本的特殊性，存在一词多义的现象。这使得一个核心动词具有多种语义，造成了根据某一动词确定的关系会出现语义上的不确定性，导致关系识别的错误。因此，在未来的科技情报对象关系抽取研究中，应更多注重非分类关系抽取和关系消歧问题，进一步提高关系抽取的多样性和准确性。

参考文献

[1] 王伟军, 王金鹏. 科学知识图谱在技术预见中的应用探析[J]. 情报科学, 2010(8):1127-1131.
[2] 李慧贞. 基于多维计量的学科结构与演化轨迹研究[D]. 郑州: 郑州大学, 2017.
[3] 田恬. 基于CiteSpace的《情报理论与实践》知识图谱计量分析[J]. 甘肃科技, 2016, 32(19):78-80.
[4] 王海燕, 冷伏海. 支持科技规划优先领域选择的战略情报与服务框架研究[J]. 图书情报工作, 2013, 57(7):70-74.
[5] 焦晓静，王兰成，韩锋．知识图谱在科技情报研究中的应用模型构建[J]．图书情报知识，2017(3):118-128.
[6] 王姣, 孙林. 基于知识图谱的2005-2015年我国情报分析研究现状[J]. 现代情报, 2016, 36(5):144-148.
[7] 李雅, 侯海燕, 朱建春,等. 知识图谱方法在科研选题中的应用研究——以乙酰甲喹纳米乳的研制及药效评价研究为例[J]. 图书情报工作, 2013, 57(9):84-91.
[8] 徐珂珂. 基于专利文献的关键技术分析方法研究[D]. 大连: 大连理工大学, 2013.
[9] 张同同. 我国科技情报研究机构科研知识图谱分析[J]. 情报探索, 2016, 1(9):124-129.
[10] Rios-Alvarado A B, Lopez-Arevalo I, Sosa-Sosa V J. Learning concept hierarchies from textual resources for ontologies construction[J]. Expert Systems with Applications, 2013, 40(15):5907-5915.
[11] 万常选, 甘丽新, 江腾蛟, 刘德喜, 刘喜平, 刘玉. 基于协陪义动词的中文隐式实体关系抽取[J]. 计算机学报, 2017, 40(76):1-24
[12] Perez-Soltero A. A semantic role labelling-based framework for learning ontologies from Spanish documents[J]. Expert Systems with Applications An International Journal, 2013, 40(6):2058-2068.
[13] Fang Q, Xu C S, Jitao Sang J T, et al. Folksonomy-based visual ontology construction and its applications[J]. IEEE Transactions on Multimedia, 2016, 18(4):702-713.
[14] Ren F. A cheap domain ontology construction method based on graph generation and conversion method[J]. Journal of Information & Computational Science, 2012, 9(18):5823-5830.
[15] Ren F. A Frequency Based Mining Method of Complex Concept Relations for Domain Ontology[J]. Journal of Information & Computational Science, 2013, 10(9):2509-2517.
[16] 曹馨宇, 曹存根, 吴昱明. 从Web 获取部分整体关系[J]. 中文信息学报, 2013, 27(2):26-33.
[17] 夏飞，曹馨宇，符建辉,等. 基于并列结构的部分整体关系获取方法[J]. 中文信息学报，2015, 29(1):88-96.
[18] Hu F, Shao Z, Ruan T. Self-supervised Chinese ontology learning from online encyclopedias[J]. TheScientific World Journal, 2014, 2014(1):848631.
[19] Ren Y, Teng C, Li F, et al. Relation classification via sequence features and bi-directional LSTMs[J]. Wuhan University Journal of Natural Sciences, 2017, 22(6):489-497.
[20] 李博, 赵翔, 王帅,等. 改进的卷积神经网络关系分类方法研究[J]. 计算机科学与探索, 2017, 10(6):1-12.
[21] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.
[22] Jarrett K, Kavukcuoglu K, Ranzato M, et al. What is the best multi-stage architecture for object recognition?[C]. IEEE, International Conference on Computer Vision. IEEE, 2010:2146 - 2153.
[23] 周超. 基于深度学习混合模型的文本分类研究[D]. 兰州: 兰州大学, 2016.