基于 BILSTM\_CＲF 的知识图谱实体抽取方法

KNOWLEDGE GＲAPH ENTITY EXTＲACTION BASED ON BILSTM\_CＲF

**（翟社平，段宏宇，李兆兆、计算机应用与软件 2019）**

**一、科学问题**

**1.1 本文所涉及科学问题**

知识图谱的实体抽取方法

**1.2 同行专家如何解决**

早期使用基于规则的方法实现实体抽取，通过领域专家和语言学者手工制定有效规则，将文本与规则进行匹配识别出实体。传统采用制定规则的实体识别方法有: Grishman开发的Proteus系统、Black开发的FACILE系统等。

知识图谱实体抽取任务被部分机器学习算法解决。常应用到实体识别任务中的模型包括隐马尔科夫模型 HMM、支持向量机模型SVM、条件随机场模型CＲF等

**1.3 本文所解决的问题**

提出一种基于 BILSTM\_CＲF 模型的神经网络结构实体抽取方法。

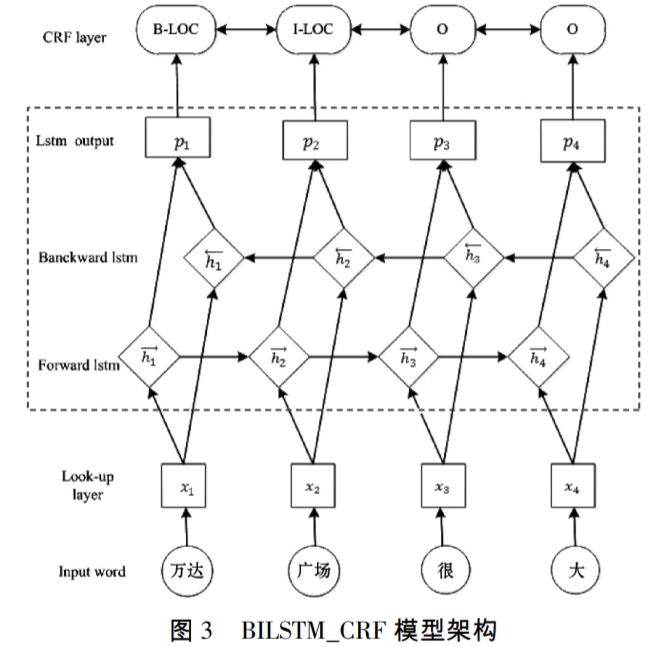
**1.4 本文解决方案效果**

基于BILSTM\_CＲF 模型的实体抽取方法有更高的准确性和查全率，其中 F1 值相比传统 CＲF 模型提高约 8%。

**二、研究内容**

**2.1 理论与方法介绍**

基于 BILSTM\_CＲF 模型的神经网络结构实体抽取方法，它既能使用双向长短时记忆网络 BILSTM( Bidirectional Long Short-Term Memory) 提取文本信息的特征，又可利用条件随机场CＲF( Conditional Random Fields) 衡量序列标注的联系。该方法对输入的文本进行建模，把句子中的每个词转换为词向量; 利用 BILSTM 处理分布式向量得到句子特征; 使用CＲF 标注并抽取实体，得到最终结果。



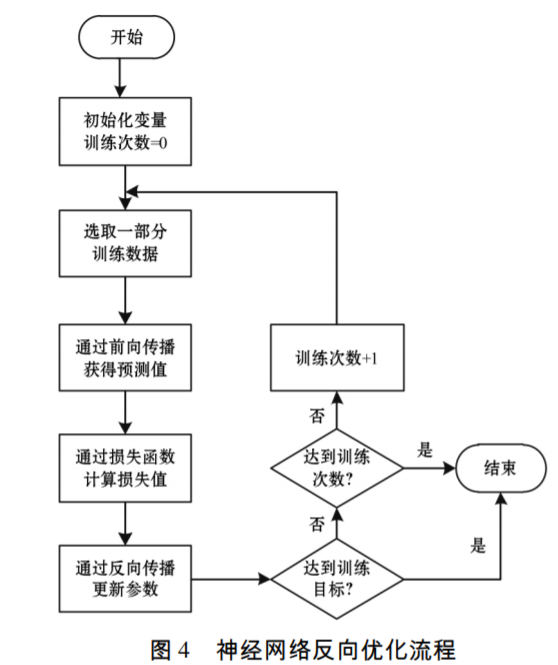
在模型架构中，look-up 是输入层；BILSTM 是隐含层；CＲF是标注层。

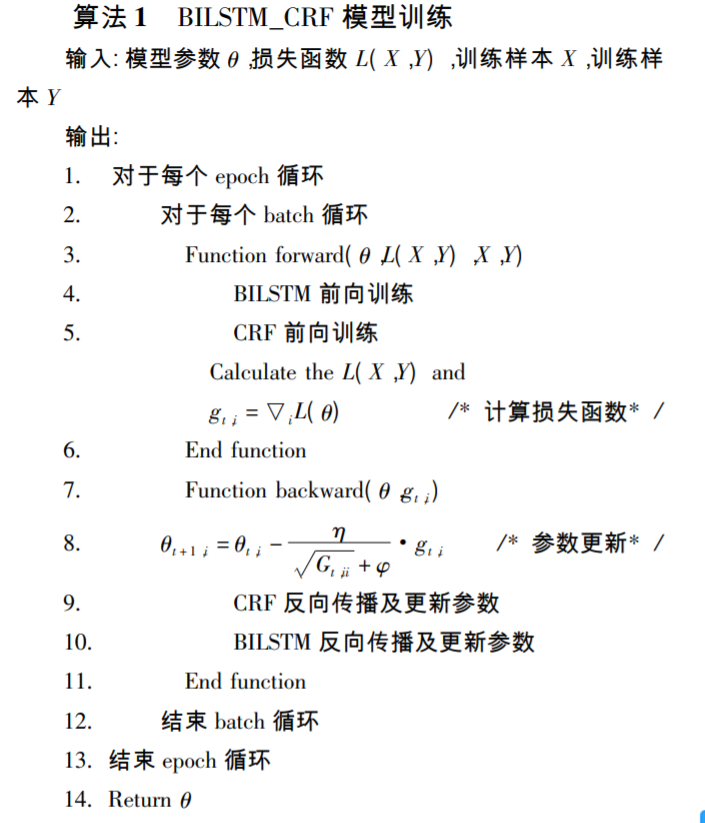
输入层主要负责将窗口的词进行向量化映射，使用训练好的词向量矩阵将输入窗口的每个词映射为分布式向量。

隐含层中，输入向量序列的顺序序列作为Banckward LSTM层的输入，逆序序列则作为Forward LSTM 层的输入。将两个序列进行位置拼接。得到完整的隐状态序列。接入一个线性变换层来实现自动获得句子特征。线性变换层把之前得到的隐状态序列从m维映射到k维，k代表标注集中所有的标签数。将结果记作矩阵 P = ( p1，p2， …，) ∈Ｒn × k ，把 pi∈Ｒk 的每一位 项都视作将字分类到第 j 个标签的打分值。

CＲF 层引入一个状态转移矩阵 M，矩阵 M 的每个元素表示从 i 变化到 j 的可能性，实现利用此前标注过的信息对一个新的位置进行标注。，对于句子 x，模型预测标 签等于 y 的打分为。最终使Softmax 函数进行归一化处理，得到结果。。

模型训练





本文引入批量样本的梯度进行参数更新，损失函数为: 式中：X，Y为批量训练样本。

参数优化利用反向传播算法可以计算损失函数对每个参数的梯度，并结合学习率更新参数，从而提高训练模型的质量。所有参数 公式为

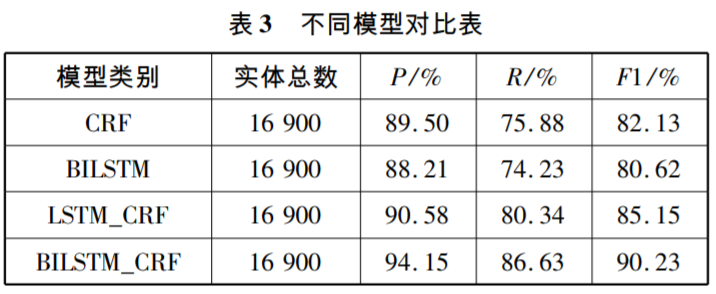
式中: φ 用来防止分母为零项，∈ 是一个对角矩阵，矩阵中的元素 表示每个参数 从最初状态到 t 时刻所有梯度平方的求和。

**2.2 验证分析与实验效果**

评价指标为准确率 P( Precision) 、召回率又叫查全率Ｒ(Ｒecall) 以及 F1 值。

、、

实验采用传统条件随机场 CＲF 模型（人工设计特征模板）、双向长短时记忆网络 BILSTM 模型，混合方法 LSTM\_CＲF 模型作为对比实验。



BILSTM 在所有测试模型中 F1 值最低。因为本实验语料库较小，模型得不到充分的训练，性能受到影响。

混合模型 LSTM\_CＲF的F1 值高，识别效果较好，但因 LSTM 网络只能提取序列标注模型的上文信息特征，无法利用下文信息特征，抽取模型未达到最好效果。

BILSTM\_CＲF 模型的实体抽取方法有更高的准确性和查全率，其中 F1 值相比传统 CＲF 模型提高约 8%。

BILSTM\_CＲF 模型的实体抽取方法的影响因素：1. 不同维度的向量。2. 受隐含层节点数目量

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

实体抽取模型训练需要大量标注语料，但在一些领域并没有海量的标注数据

**3.2 后续研究重点**

以基于BILSTM\_CＲF 的知识图谱实体抽取方法的所提方法引入迁移学习，使用少量的标注数据进行实体抽取将 是以后研究的重点。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：基于 CＲF 和规则相结合的地理命名实体识别方法

**（2）作者介绍**：何炎祥12、罗楚威2、胡彬尧2，1.武汉大学软件工程国家重点实验室、2.武汉大学计算机学院

**（3）摘要**: 为了识别文本中海量的地理命名信息，以 CＲF( 条件随机场) 模型识别为基础，加入制定的规则，来提高 CＲF 模型识别的召回率，从而提高整体的地理命名实体识别效果。通过选取适合的地理命名实体识别的特征模板，验证特征的有效性以及分析 CＲF模型识别结果中的未识别实体样本，设计针对未识别实体的规则用以修正识别结果。实验表明，对地名和组织名结合规则进行修正后的 F 值达到了 91． 61% 和 85． 74% ，有了显著提高。

**4.2 相关论文二**

**（1）题目**：基于BERT嵌入的中文命名实体识别方法

**（2）作者介绍**：杨飘、董文永，武汉大学计算机学院

**（3）摘要**：命名实体识别是自然语言处理的基础之一。在基于神经网络的中文命名实体识别方法中,字的向量化表示是重要一步,传统的词向量表示将字映射为单一向量,这种方法无法表征字的多义性。针对这个问题,提出了基于BERT嵌入的中文命名实体识别方法。该方法通过BERT(Bidirectional Encoder Representations from Transformers)预训练语言模型增强字的语义表示,根据字的上下文动态生成语义向量,然后再将字向量序列输入BiGRU-CRF中进行训练,训练时可以训练整个BERT-BiGRU-CRF模型,也可以固定BERT,只训练BiGRU-CRF部分。实验表明,该模型的两种训练方法在MSRA语料上分别达到95.43%F1值和94.18%F1值,均优于目前最优的Lattice-LSTM模型。

**4.3 相关论文三**

**（1）题目**：基于SVM-BiLSTM-CRF模型的财产纠纷命名实体识别方法

**（2）作者介绍**： 周晓磊2、赵薛蛟12、刘堂亮3、宗子潇4、王其乐5、里剑桥6

1.中国科学院大学 2 .中国科学院沈阳计算技术研究所3 .辽宁省人民检察院沈阳铁路运输分院4.东北大学 5 .沈阳市第三十一中学6 .大连理工大学

**（3）摘要**：裁判文书中的命名实体识别是自动化审判的关键一步,如何能够有效的分辨出案件的关键命名实体是本文的研究重点.因此本文针对财产纠纷审判案件,提出了一种基于SVM-BiLSTM-CRF的神经网络模型.首先利用SVM筛选出包含关键命名实体的句子,然后将正确包含此类实体的句子转化为字符级向量作为输入,构建适合财产纠纷裁判文书命名实体识别的BiLSTM-CRF深层神经网络模型.通过构建训练数据进行验证和对比,该模型比其他相关模型表现出更高的召回率和准确率。

**知识图谱技术综述**

**Review on Knowledge Graph Techniques**

**徐增林，盛泳潘,贺丽荣，王雅芳、电子科技大学学报**

**一、科学问题**

**1.1 本文所涉及科学问题**

本论文主要是讨论知识图谱技术，在全面阐述知识图谱定义、架构的基础上，综述知识图谱中的知识抽取、知识表示、知识融合、知识推理四大核心技术的研究进展以及一些典型应用。

**1.2 知识图谱的现状**

知识图谱是在2012年由Google正式提出，它被广泛应用于智能搜索、智能问答、个性化推荐等领域知识图谱能够使计算机理解人类的语言交流模式，从而更加智能地反馈用户需要的答案。与此同时，通过知识图谱能够将Web上的信息、数据以及链接关系聚集为知识，使信息资源更易于计算、理解以及评价，并且形成一套Web语义知识库。

**1.3 本文的结构**

第一部分对知识图谱的定义与架构进行解析；第二部分简要介绍了开放连接知识库和垂直行业知识库；第三部分对知识图谱的知识获取、知识表示、知识融合和知识推理的技术研究和技术细节进行阐述；第四部分介绍知识图谱在智能搜索、深度问答、社交网络以及垂直行业的应用情况；第五部分知识图谱所面临的一些困难与挑战。

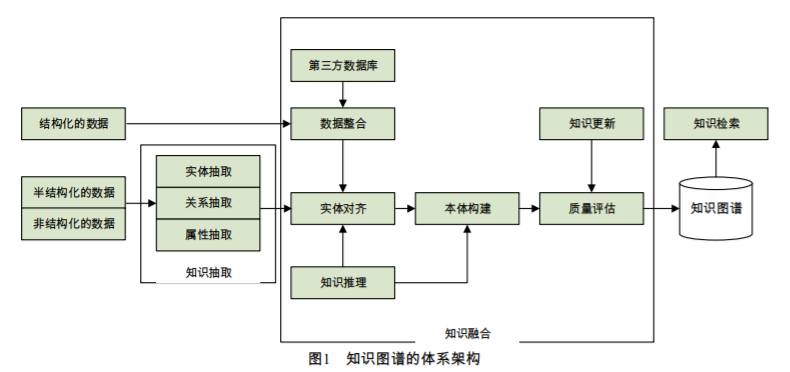
**二、论文内容**

第一部分：知识图谱的定义与架构

知识图谱的定义是Google用于增强其搜索引擎功能的知识库，也是一种揭示实体之间关系的语义网络。

知识图谱的通用表达方式是三元组，即G=（E，R，S）：E为知识库的实体集合；R为知识库的关系集合；S ⊆E×R× E代表知识库中的三元组集合。

知识图谱的架构主要包括自身的逻辑结构以及体系架构。逻辑结构分为模式层与数据层；体系架构有自顶向下和自底向上（大多数）两种。



第二部分

开放链接知识库：Freebase、Wikidata、DBpedia、YAGO，其数据大多数是半结构化和非结构化的。

垂直行业知识库：MusicBrainz、IMDB、豆瓣

第三部分：知识图谱的关键技术：

1知识抽取

实体抽取其方法有三种：基于规则与词典的方法（覆盖领域有限，很难适应数据变化的需求）、基于统计机器学习的方法（算法的准确率与召回率不理想）以及面向开放域的抽取方法。

关系抽取的抽取方法：早期使用人工构造语义规则以及模板的方法识别实体关系；使用实体间的关系模型（需提前定义实体间的关系类型）；面向开放域的信息抽取框架（性能低下）；基于马尔可夫逻辑网、基于 本体推理的深层隐含关系抽取方法。

属性抽取：可转换为关系抽取；基于规则与启发式算法的属性抽取方法；从百科网站上抽取的结构化数据作为可用于属性抽取的训练集，然后再将该模型应用于开放域中的实体属性抽取；是根据实体属性与属性值之间的关系模式，直接从开放域数据集上抽取属性（准确率不高）。

2.知识表示

基于三元组的表示形式虽说得到了广泛的认可，但计算效率，数据稀疏性却存在很多问题。知识表示学习的代表模型、复杂关系翻译模型、多源异质信息融合模型的出现。

代表模型主要包括距离模型（协同性差）、 双线性模型（形式简单，易于计算）：衍生出DISTMULT模型（实验效果好）、神经张量模型、单层神经网络模型（计算开销大）、矩阵分解模型、翻译模型（计算复杂度低）等。

复杂关系模型：TransH模型、TransR模型、TransD模型、TransG模型、KG2E模型

3.知识融合

实体对齐：1) 将待对齐数据进行分区索引，以降低计算的复杂 度；2) 利用相似度函数或相似性算法查找匹配实 例；3) 使用实体对齐算法进行实例融合；4) 将步骤 2)与步骤3)的结果结合起来，形成最终的对齐结果。对齐算法可分为成对实体对齐与集体实体对齐两大类，而集体实体对齐又可分为局部集体实体对齐与全局集体实体对齐。

知识加工：本体构建（人工编辑、数据驱动自动构建）、质量评估

知识更新：逻辑结构的模式层和数据层的更新

知识推理：知识推理方法主要可分为基于逻辑的推理、基于图的推理、基于组合描述逻辑的Tableau算法。基于逻辑的推理要包括一阶谓词逻辑 (first order logic)、描述逻辑(description logic)以及规则等；

第四部分：知识图谱的典型应用

智能搜索：1) 查询式语义理解，2) 知识检索。展现知识的形式有1) 集成的语义数据（知识卡片配图片等）。2) 直接给出用户查询问题的答案。3) 根据用户的查询给出推荐列表

深度问答：是信息检索系统的一种高级形式（查询式理解与知识检索）

社交网络：就是通过知识图谱将人、 地点、事情等联系在一起，并以直观的方式支持精确的自然语言查询

垂直行业（如金融、医疗、电商……）

第五部分：知识图谱的挑战

知识抽取：由于制约条件较多，方法的可扩展能力不够强，未能很好地适应大规模、领域独立、高效的开放式信息抽取要求。对于隐含关系的抽取，目前主流的开放式信息抽取方法性能低下或尚无法实现。

知识表示：基于三元组形式完成的语义映射在面对复杂的知识类型、多源融合的信息时，其表达能力仍然有限。研究方向：对认知科学领域人类知识类型的探索将有助于知识类型的划分、表示以及处理，是未来知识表示研究的重要发展方向；基于实体的、关系的、Web文本的、多知识库的融合均具有较为广阔的研究前景

知识融合：大规模的知识库不仅蕴含了海量的知识，其结构、数据特征也极其复杂，这些对知识库实体对齐算法的准确率、执行效率提出了一定的挑战。

将众包平台与知识库对齐模型有机结合起来，并且能够有效判别人 工标注的质量，这些均具有较为广阔的研究前景。

多语言知识库对齐的质量不高，这方面仍有广阔的研究空间。本体自动构建、本体抽取、本体聚类等问题是目前的研究热点。现有的更新技术均过多依赖于人工的干预，增量更新技术将是知识图谱未来实现自动化更新的重要研 究方向。如何确保自动化更新的有效性，是更新过程中面临的又一重大挑战。

智能搜索、深度问答、社交网络以及其他行业中的使用也只是处于初级阶段，仍具有广阔的可扩展空间。

**三、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：知识图谱构建技术综述

**（2）作者介绍**：刘峤、李杨、段宏、刘瑶、 秦志光 ，电子科技大学信息与软件工程学院

**（3）摘要**: 谷歌知识图谱技术近年来引起了广泛关注，由于公开披露的技术资料较少，使人一时难以看清 该技术的内涵和价值．从知识图谱的定义和技术架构出发，对构建知识图谱涉及的关键技术进行了自底向上的全面解析．１）对知识图谱的定义和内涵进行了说明，并给出了构建知识图谱的技术框架，按照输 入的知识素材的抽象程度将其划分为３个层次：信息抽取层、知识融合层和知识加工层；２）分别对每个层次涉及的关键技术的研究现状进行分类说明，逐步揭示知识图谱技术的奥秘，及其与相关学科领域的关系；３）对知识图谱构建技术当前面临的重大挑战和关键问题进行了总结。

**4.2 相关论文二**

**（1）题目**：知识图谱研究综述

**（2）作者介绍**：黄恒琪1 ,  于娟1 ,  廖晓2 ,  席运江3

1.福州大学 经济与管理学院2 .广东金融学院 互联网金融与信息工程学院 3 .华南理工大学 工商管理学院

**（3）摘要**：知识图谱是以图的形式表现客观世界中的概念和实体及其之间关系的知识库, 是语义搜索、智能问答、决策支持等智能服务的基础技术之一. 目前, 知识图谱的内涵还不够清晰; 且因建档不全, 已有知识图谱的使用率和重用率不高. 为此, 本文给出知识图谱的定义, 辨析其与本体等相关概念的关系. 本体是知识图谱的模式层和逻辑基础, 知识图谱是本体的实例化; 本体研究成果可以作为知识图谱研究的基础, 促进知识图谱的更快发展和更广应用.本文罗列分析了国内外已有的主要通用知识图谱和行业知识图谱及其构建、存储及检索方法, 以提高其使用率和重用率. 最后指出知识图谱未来的研究方向。

**4.3 相关论文三**

**（1）题目**：知识图谱研究综述

**（2）作者介绍**：李涓子、侯磊 , 清华大学 计算机科学与技术系

**（3）摘要**：知识图谱以结构化的方式描述客观世界中概念、实体及其间的关系，将互联网的信息表达成更接近人类认 知世界的形式，提供了一种更好地组织、管理和理解互联网海量信息的能力。知识图谱与大数据和深度学习一起， 正在成为推动人工智能发展的核心驱动力。文章概述了当前已有的重要知识资源，对知识图谱关键技术—知识表示、知识图谱构建和知识图谱应用进行了综述，并对知识图谱未来发展的挑战和趋势进行了总结展望。