基于改进 BERT 算法的专利实体抽取研究——以石墨烯为例

李 建[1]，靖富营[2]，刘 军[1*]

(1. 电子科技大学经济与管理学院 成都 610054；2. 重庆工商大学国家智能制造服务国际科技合作基地 重庆 南岸区 400067)

【摘要】实体关系抽取是判断专利新颖性的核心环节，传统的实体关系抽取都是采用串行方式来进行，有很大的局限性。该文利用两种改进的 BERT 算法研究了专利实体关系抽取的技术演化。一种是将中文特征和句法语义特征相结合的新算法——基于改进的 BERT-BiLSTM-CRF 命名实体识别算法；另一种是将注意力机制与句法语义特征相结合的新算法——基于注意力机制与语义结合的实体关系抽取算法。最后以石墨烯制备技术为例，利用数值实验说明改进的两种算法能够高效分析专利的内容，揭示石墨烯企业技术的动态演化过程。

关 键 词 演化分析； 实体抽取； 石墨烯技术； 专利

中图分类号 TP312；F274 **文献标志码** A doi:10.12178/1001-0548.2020132

Study on Patent Entity Extraction Based on Improved Bert Algorithms——A Case Study of Graphene

LI Jian[1], JING Fu-ying[2], and LIU Jun[1*]

(1. School of Management and Economics, University of Electronic Science and Technology of China Chengdu 610054;
2. National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University Nanan Chongqing 400067)

Abstract The entity relation extraction is the key part to estimate the novelty of patents. The traditional entity relation extraction is the series system, but this style has major dwawbacks. The paper studies the evolution of entity relation extraction using two improved BERT algorithms. One is the method combining traditional Chinese features with syntactic semantic features, and the other is the method combining attention mechanism with syntactic semantic features. The extensive computational experiments and the preparation technology of the graphene show that the two algorithms can improve the analysis efficiency for the contents of the patents and reveal the dynamic evolution process of the technology of the graphene firm.

Key words evolutionary analysis; entity extraction; graphene technology; patent

专利文档中含有大量作者所进行的创新性工作，这些内容所蕴含的知识代表先进技术，对专利文档的分析可以获得专利所研究领域的技术及生产工艺发展情况。但是由于专利文档数量的庞大性，如果每一篇都需要人工分析和信息提取的话，则工作量非常大，同时也会受到操作者本身技术能力的影响，因此采用自动获取技术是专利分析的第一要素。自然语言处理在近些年来成功应用到诸多文档处理相关领域，获得了显著效果。基于实体关系的知识图谱技术也是采用符合人类社会模型认知的方式来深入挖掘实际事物之间的联系，进而完成知识演进。专利文本中核心的文档主要是说明书和权利要求，这两部分包含了专利的大多数信息，权利要求以科学术语定义该专利或专利申请所给予的保护范围。说明书则是对发明或者实用新型的结构、技术要点、使用方法做出清晰、完整的介绍，它包含了背景技术、发明内容、附图说明、具体实施方案等项目。本文将采用两种算法进行专利信息的抽取，实现对专利文本中的核心涉及物及关键工艺的认知。

在专利知识抽取方面，国内有学者探索了基于规则、模板、机器学习、本体等多种抽取的方法。

收稿日期：2020－03－23；修回日期：2020－07－01
基金项目：国家自然科学基金 (71874023)
作者简介：李建 (1986－)，男，博士，主要从事产业创新政策方面的研究.
通信作者：刘军，E-mail：lj@uestc.edu.cn

文献 [1] 研究了专利摘要信息抽取的技术、步骤，结合词典、规则和统计模型方法，针对隐马尔可夫标注算法进行了合理改进，在抽取结果处理上提出了一套技术关键词识别模型及其算法。文献 [2] 提出了针对英文专利的，基于模板的自动获取方法。文献 [3] 提出一个基于本体的中文专利摘要抽取模型。文献 [4] 在领域专利术语抽取的基础上，研究较大规模术语层次关系的解析，构建了含有层次关系的领域知识本体。文献 [5] 研究了使用不完备的语料库，在无人工参与的情况下，采用条件随机场的方法对字进行角色的标注，并设计术语识别的模型，取得了较好的效果，从专利中抽取的知识可用于辅助技术或产品创新。文献 [6] 研究了基于同义词群提取的技术特征，用于外观设计专利的分析。国外在专利标注和知识抽取方面也有研究，文献 [7] 根据专利文档的结构和语义描述，对专利进行语义标注，帮助生物学家更好的利用专利信息。文献 [8] 基于文档结构以及专利文档内容的语义结构，利用自然语言和本体技术，对专利进行语义标注，便于对专利检索更好的分析。文中还描述了专利分析人员分析过程中用到的一系列文本挖掘技术方案，包括文本切分、摘要抽取、特征项选择、词语关联、聚类、主题识别和信息映射等。结果证明自动抽取的概要相比其他片段更能表达原来的意思。这些技术有助于提高专利分析中用到的分类、组织、知识分享和现有技术检索。文献 [9] 提出了一种基于语义要素统计和关键短语抽取的中文专利挖掘方法，用于从中文专利文档中抽取关键短语。该抽取技术基于“HowNet”的语义知识结构，利用统计的方法计算专利文档中的备选短语计算值。实验证明，该方法比单纯的频次统计方法有更高的精确率和召回率。文献 [10-11] 介绍了一种词间关系抽取的方法，结合模板和统计指标来抽取词间的两种类型的层次关系：“IS-A”和“PART-OF”。

1 经典实体信息抽取技术

1.1 专利实体及实体关系内容

专利实体内容指专利中所应用到的实体，包括化学材料、实验器材等，例如氧化石墨烯、碳纤维、烘箱、真空泵。这些实体都对整个专利的制作流程起着重要作用，而承载着这些实体的就是操作，每篇专利里都会对各种实体进行不同的操作，以此达到不同的目的，同种材料的不同操作方式、不同操作顺序也是一篇专利的创新性和新颖性的体现。对专利的实体内容进行抽取分析，对比各个专利所使用的材料差异和操作差异，再结合对应的评价体系，最终得到专利的创新性和新颖性评价指标。

实体关系抽取是判断专利新颖性的核心环节，其任务是从大量专利文本数据中抽取出能够表达专利工艺流程的结构化动宾关系，也就是关系双元组。例如以石墨烯制备技术为例，在专利说明书中的发明内容中的工艺流程：<得到，氧化石墨烯分散液>，利用基于字级别的字符串搜索技术在专利中检索到包含此流程的原句为：“将氧化石墨烯 (GO) 于水中分散，得到氧化石墨烯分散液”。从这句话中得出抽取出来的“得到氧化石墨烯分散液”为此专利工艺流程的其中一步，以此种方法为例，最终可得到整个专利发明内容内的实体关系列表，整个列表又可作为专利的工艺流程序列，最终得到此专利的技术方案。

1.2 传统信息抽取算法

文本文档是典型的非结构化信息，不能像数据库之类的信息可以通过键值对来进行数据分析和统计，但是文档却不限制文本的结构内容，进而可以承载更多的信息。非结构化文档信息抽取技术就是通过自然语言分析技术来实现对其核心内容的信息获取，其中实体与关系抽取是目前最为成功的技术。目前传统的实体抽取和关系抽取都是采用串行的方式来进行，先完成对实体的识别与提取，然后再分析不同实体之间的关系。

实体抽取第一步是进行命名实体识别 (named entity recognition, NER)，目前通用文档中识别的命名实体主要是人物 (person, PER)、地点 (location, LOC)、机构 (organization, ORG)、时间 (time, TIME)、数字 (number, NUM)、描述 (description, DES) 和混杂 (miscellaneous, MISC)。但是考虑到所需处理的任务，就需要对所识别的实体进行调整。非结构化文档的实体与关系抽取的传统流程为：首先对输入文本进行预处理，预处理主要完成分词、停用词处理和词性标注，获得比较纯粹的文本词语。然后将处理之后的文本输入命名实体识别模型中，在该模块中主要完成对命名实体的识别，一般是采用从前到后的处理过程，根据前后关系和句法分析等方式来对输入的词汇进行判断。在此过程中，词汇的前后顺序也是非常关键的信息。命名识别完成之后就实现了文本的序列标注，该结果可以输出到实体集识别内，也是下一步实体关系识别的输入。关系识别需要对输入的多实体和其顺序标注进行处理，通

过学习关系模型，可以获得模型可识别的关系，比如位置关系、工作关系、隶属关系等，流程如图 1 所示。

针对石墨烯专利文本，本文采用图 2 中传统框架完成对专利文本信息的抽取识别。

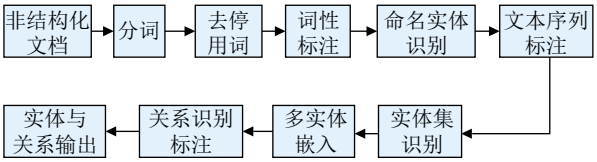

图 1　传统实体关系抽取流程

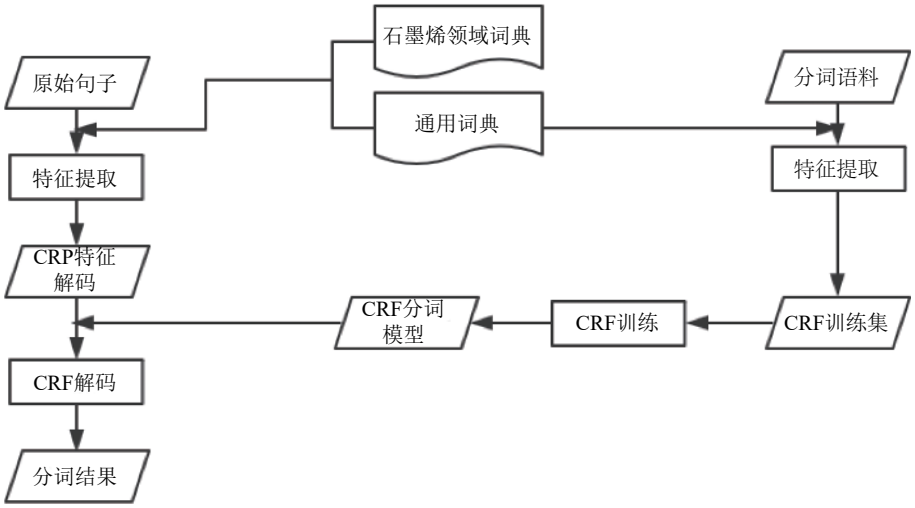

图 2　基于传统算法的专利中文实体关系抽取框架图

该框架是基于最为流行的 CRF 算法模型为基础来进行的。这种方式提升了在预处理阶段分词的准确率，使其可以更精准地识别专利中的词汇。输入专利文档中的语句，通过切词、词性标注以及依存句法分析，由于专利分析中不需要依赖特定词语，所以在此框架中采用类似最大匹配的思想，将目前包含最长字符词的长度信息提供给统计模型。

句法分析是处理过程中重要的一环，该过程需要精准的完成句子内各词汇之间的依存关系分析，进而可以分析语句的构成和依赖关系。依赖关系分析通过分解句子各词语之间的语义关系来刻画句子语意，并且将语义层面的关系用依赖结构模式展现，其不用对词语进行抽象表达，直接利用词语所处的语义关系结构来表征词汇。句法结构的解析也就是短语结构解析，即分析句子中词汇间的相互联系、相互作用的方式。该框架首先对句子进行分词和词性标注，再进行句法结构分析，将分析结果以句法分析树的结构展示，从而进一步明确每个词汇以及短语在句子结构中承担的作用，帮助识别最有可能与关系特征词构成关系的实体对。本框架所分析句法依存关系如表 1 所示。

表 1　句法依存关系

关系	标签	描述
主谓关系	SBV	subject-verb
动宾关系	VOB	verb-object
间宾关系	IOB	indirect-object
前置宾语	FOB	fronting-object
兼语	DBL	double
定中关系	ATT	attribute
状中结构	ADV	adverbial
动补结构	CMP	complement
并列关系	COO	coordinate
介宾关系	POB	preposition-object
左附加关系	LAD	left-adjunct
右附加关系	RAD	right-adjunct
独立结构	IS	independent-structure
核心关系	HED	head

2　BERT 专利实体关系抽取算法

2.1　BERT 模型

BERT 是一种新型的语言模型，使用 Transformer 做 encoder，可以用更深的层数，具有更好的并行性，它通过联合调节所有层中的双向 transformer 来训练预训练深度双向表示。BERT 基于所有层中的左、右语境进行联合调整，来预训练深层双向表征，因此，只需要增加一个输出层，就可以对预训练的 BERT 表征进行微调，为更多的任务创建当前

的最优模型，如本节中要解决的实体关系抽取任务。BERT 模型结构如图 3 所示。

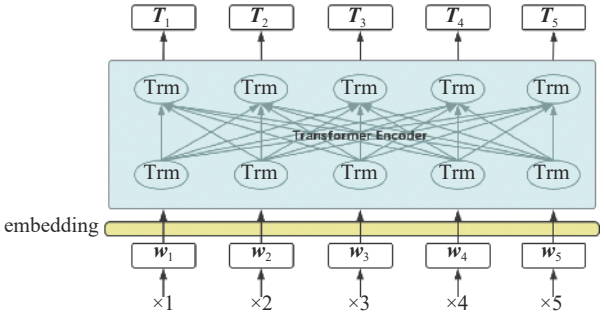

图 3　BERT 模型结构

图中，w_i 表示由每个参数 x_i 对应转化来的词向量，本文采用 embedding 的向量方式。embedding 层由 3 种 embedding 求和而成；在 embedding 后将组合向量进行编解码运算 (Trm 层)，之后将每个单词的特征向量 (T) 作为结果输出。在 embedding 计算中，第一个单词是 CLS 标志，可以用于之后的分类任务，为区别两个句子，用一个特殊标志 SEP 隔开它们，另外针对不同的句子，把学习到的 segment embeddings加到每个 token 的 embedding 上面；segment embeddings 用来区别两种句子，因为预训练不只做 LM (language model) 还要做以两个句子为输入的分类任务，position embeddings 是学习结果，其结构如图 4 所示。

将 BERT 模型应用到专利文本信息抽取上，需要采用如图 5 所示的模型。

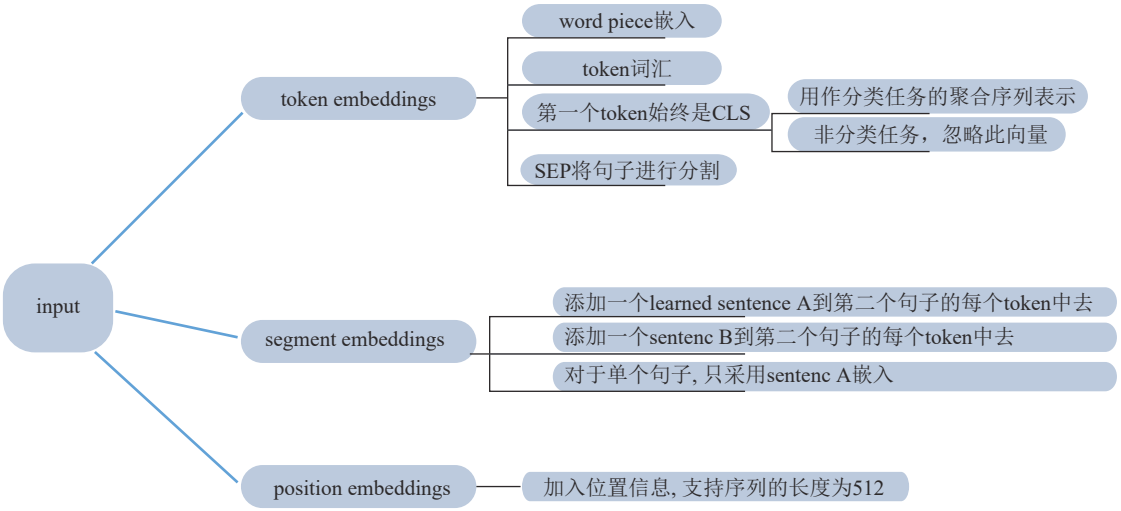

图 4　BERT 输入表征示意图

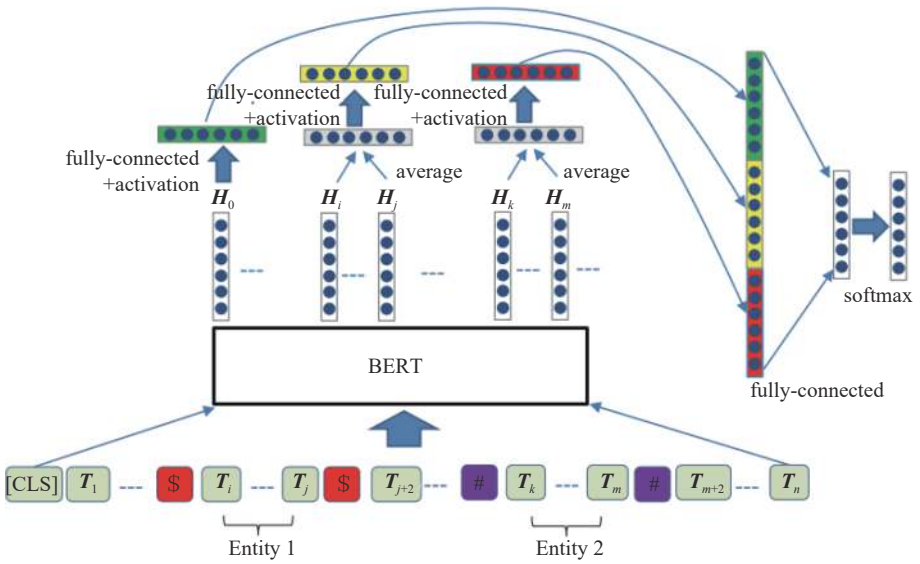

图 5　专利抽取 BERT 模型体系结构

此模型的核心步骤分为：

1) 目标实体由基础的语言单元字或词组成 (图中 $\boldsymbol{T}_k$)，为了能够定位两个目标实体并将其信息转移到 BERT 中，在将整个问题投入 BERT 前，在目标实体前后添加 token，即符号“$”和“#”；

2) 通过 BERT 输出目标实体对应的输出进行定位；

3) 利用 BERT 输出的 [CLS] 隐含向量和两个目标实体的隐含向量进行关系分类。

假设输入的句子为“四川的省会是成都”，将此句子输入进模型的第一层，将会在它的开头添加 [CLS] 符号，该句子为单句，不需要添加 [SEP]，BERT 模型的输出部分包括 3 个部分，第一部分为 [CLS] 标签，第二部分为第一个实体的隐含向量，第三部分为第二个实体的隐含向量。这样第一部分可以保存整个句子的语义内容，后两部分则是保存实体的信息。

再将识别到的第一个实体前后加入$符号—“[CLS]$四川$的省会是成都”，最后将识别到的第二个实体前后添加#符号—“[CLS]$四川$的省会是#成都#”，两个实体前后添加特殊符号的目的是标识两个实体，让模型能够知道这两个词的特殊性，相当于变相指出两个实体的位置。此时输入的维度为 [batch size n, max_length m, hidden size d]。[CLS] 位置的输出可以作为句子的向量表示，记作 $\boldsymbol{H}_0$，维度是$[n,d]$，经过线性变换后添加 tanh 激活函数得到，$\boldsymbol{W}_0$的维度是$[n,d]$，因此$\boldsymbol{H}_0'$的维度就是$[n,d]$。$\boldsymbol{b}_0$ 是实体偏移量，由每个实体在句子中的前后位置决定。[CLS] 表征：该部分为单一向量，因此直接将其输入前馈神经网络中，可表示为：

$$\boldsymbol{H}_0' = \boldsymbol{W}_0(\tanh(\boldsymbol{H}_0)) + \boldsymbol{b}_0 \tag{1}$$

除了利用句向量之外，模型还结合了两个实体向量。实体向量通过计算 BERT 输出的实体各个字向量的平均得到，假设 BERT 输出的实体 1 的开始和终止向量为$\boldsymbol{H}_i$，$\boldsymbol{H}_j$。实体 2 为$\boldsymbol{H}_k$，$\boldsymbol{H}_m$。其中i、j、k、m分别为第一个实体的首字符位置、第一个实体的末字符位置、第二个实体的首字符位置、第二个实体的末字符位置。那么实体 1 和 2 的向量表示为：

$$\boldsymbol{e}_1 = \frac{1}{j-i+1}\sum_{t=i}^{j}\boldsymbol{H}_t \tag{2}$$

$$\boldsymbol{e}_2 = \frac{1}{m-k+1}\sum_{t=k}^{m}\boldsymbol{H}_t \tag{3}$$

得到的实体向量也需要经过激活函数和线性层，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$的维度都是$[d,d]$，实体信息为：

$$\boldsymbol{H}_1' = \boldsymbol{W}_1\boldsymbol{e}_1 + \boldsymbol{b}_1 \tag{4}$$

$$\boldsymbol{H}_2' = \boldsymbol{W}_2\boldsymbol{e}_2 + \boldsymbol{b}_2 \tag{5}$$

最后把$\boldsymbol{H}_0', \boldsymbol{H}_1', \boldsymbol{H}_2'$连接起来得到一个综合向量$[n,3d]$维，输入到线性层并做 softmax分类，其中$\boldsymbol{W}_3$的维度是[关系数量 L,$3d$]，因此$\boldsymbol{h}''$的维度是$[n,L]$，得到每句话的关系类别概率分布为：

$$\boldsymbol{h}'' = \boldsymbol{W}_3[\mathrm{concat}(\boldsymbol{H}_1', \boldsymbol{H}_1', \boldsymbol{H}_2')] + \boldsymbol{b}_3 \tag{6}$$

$$p = \mathrm{soft\,max}(\boldsymbol{h}'') \tag{7}$$

使用本文改进的 BERT 模型对专利工艺流程进行提取，提取到的结果噪声较小，具体效果如图 6 所示。

动宾　(阻燃，纳米复合材料制备及其方法)
动宾　(附着，氧化石墨烯)
动宾　(提供，氢氧化镁)
动宾　(提供，石墨烯复合材料的)
动宾　(混合，均匀特征时间)
动宾　(提供，氢氧化镁石墨烯复合材料)
动宾　(制备，氧化石墨烯)
动宾　(得到，氧化石墨烯分散液)
动宾　(配制，镁盐溶液)
动宾　(配制，碱溶液)
动宾　(加入，旋转床或套管)
动宾　(加入，镁盐溶液)
动宾　(加入，碱溶液)
动宾　(得到，氢氧化镁)
动宾　(加入，镁盐溶液)
动宾　(加入，碱溶液)
动宾　(加入，氧化石墨烯表面负电荷)
动宾　(加入，高锰酸钾)
动宾　(升至，一定温度)
动宾　(得到，一定的氧化石墨烯分散液)
动宾　(缩短，30 s~30 min)
动宾　(提供，有力保障)
动宾　(加入，石墨烯分散液)
动宾　(加入，镁盐溶液)
动宾　(加入，碱溶液)
动宾　(混合，石墨烯溶液镁盐溶液以及碱液)

图 6　专利抽取 BERT 模型体系结构

从图中可以看出，使用 BERT 进行关系抽取后，整个专利的工艺流程更加清晰合理，去掉了基于 CRF 和依存句法树的实体关系抽取方法中识别不到的噪声，使结果更加精简且符合逻辑，但是从上图可以看出，虽然此方法比基于 CRF 和句法依存树的效果更好，但噪声仍然对识别流程产生较大的影响。

2.2　BERT-BiLSTM-CRF 命名实体识别算法

双向长度记忆网络采用 3 层结构来实现最终的抽取功能，字符嵌入层用以完成输入的字符向网络的输入的转换，使其利于后期的识别。Bi-LSTM 层利用长短记忆网络来实现文本在序列空间上的关联

分析，进而发现合适有价值的命名实体。BiLSTM-CRF 模型是将双向长短记忆网络和 CRF 模型结合起来，即在双向长短记忆网络的 Hide 层后再加一层 CRF，模型结构如图 7 所示。

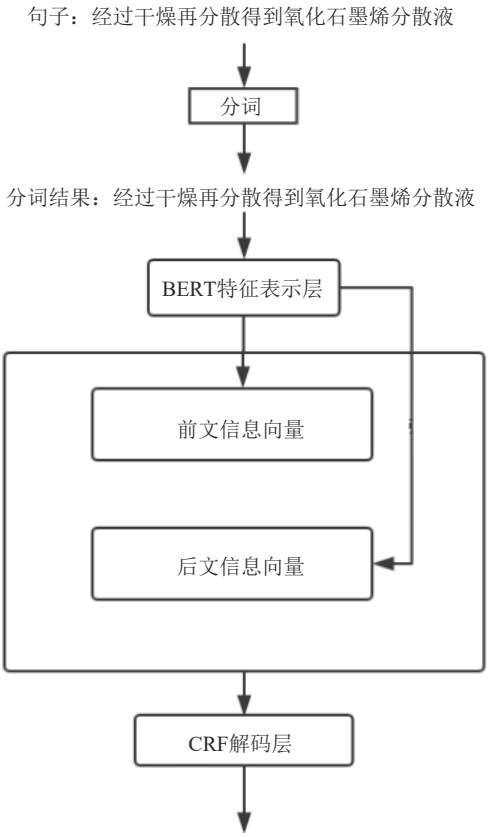

图 7 BERT-BiLSTM-CRF 命名实体识别模型结构图

由图 7 可以看出，此模型是由 BERT 产出的特征作为输入到双向长短记忆网络中，再通过 CRF 进行解码。

2.3 实体关系抽取算法

当前关于实体关系抽取的实现一般还是基于统计学方法与模式匹配，如前面提到的，先用 NLP 处理工具对词语和句子进行提取分析，再使用词与词之间的联系进行句法分析，最后抽取出目标实体和关系。传统的关系抽取方法效果虽然不错，但是它不仅需要花费大量的人工成本，而且容易出现依存关系不全以及关系混乱等错误，影响后续实验结果。除此之外，使用 NLP 处理工具提取出来的词汇信息往往会出错，如将名词动词识别错误，这些错误特征会对实体关系抽取的结果产生不利影响。近年来，深度神经网络在自然语言处理领域已经取得了许多突破性的成果，使得越来越多的研究者开始关注将深度学习与实体关系抽取结合的应用。

本节提出了一种基于注意力机制融合句子语义的实体关系抽取方法。结合汉语语言的特性，通过句子语义特征，分析句子语义相似性，构建句子特征表示并输入到模型，如图 8 所示。

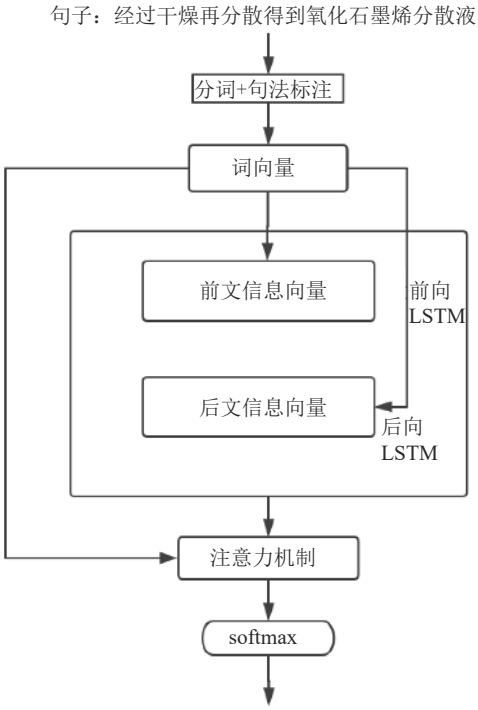

图 8 注意力机制融合句子语义的模型结构

由图 8 所示的模型结构，输入一个句子，首先将句子进行分词，再输入进 BERT 标注后，将训练后的分布式词向量输入到 BiLSTM 中获取句子语义信息，引入字级别的注意力机制来关注句子中的重要信息，通过注意力机制赋给字权重，自动获取对实体关系抽取有较大影响力的字节，最后输入权重向量通过 softmax 对关系进行分类处理，最终结果如图 9 所示。

动宾 （提供，氢氧化镁石墨烯复合材料）
动宾 （制备，氧化石墨烯）
动宾 （得到，氧化石墨烯分散液）
动宾 （配制，镁盐溶液）
动宾 （配制，碱溶液）
动宾 （加入，旋转床或套管）
动宾 （加入，镁盐溶液）
动宾 （加入，碱溶液）
动宾 （得到，粗产物）
动宾 （得到，氢氧化镁）
动宾 （加入，氧化石墨烯表面负电荷）
动宾 （加入，高锰酸钾）
动宾 （升至，一定温度）
动宾 （得到，一定的氧化石墨烯分散液）

图 9 注意力机制融合句子语义的实体关系抽取结果

2.4 算法结果分析

针对本文提及的基于改进的 BERT-BiLSTM-

CRF 命名实体识别和注意力与句法结合的实体关系识别所形成的整体框架，本文采用精确率 (P)、召回率 (R) 和$F1$值来进行上述算法的抽取效果，定义为：

$$P = \frac{N_r}{N_p} \tag{8}$$

$$R = \frac{N_r}{N_s} \tag{9}$$

$$F1 = \frac{2PR}{P+R} \tag{10}$$

式中，N_r表示预测正确的动宾关系的句子数目；N_p表示待预测的句子数目；N_s表示标准动宾结果的句子数目。

表 2 中汇总了本文中各模型的实验结果，从中可以看出无论从精确率、召回率和 $F1$ 值来看，基于改进的 BERT-BiLSTM-CRF 命名实体识别和注意力与句法结合的实体关系识别算法效果都要优于 CRF+依存句法和单纯的 BERT 算法。具体来讲，CRF+依存句法计算出的精确率、召回率和 $F1$ 值分别为改进的 BERT-BiLSTM-CRF 算法效果的 39.8%，47.2% 和 43.3%；BERT 算法的精确率、召回率和 $F1$ 值为改进的 BERT-BiLSTM- CRF 算法效果的 73.7%。

表 2　模型实验结果

%

Model	实验结果		
	P	R	$F1$
CRF+依存句法	25.18	27.84	26.44
BERT	46.67	43.43	44.99
BERT+BiLSTM+CRF结合注意力与句法关系	63.33	58.89	61.03

3　结 束 语

专利内容的创新性和新颖性是研究专利在常规普适性内容和突破性研究的评价基础，构建专利知识评价网络来为评估其内容新颖性和对后期该领域的发展所形成的突破性贡献奠定基础。本文重点描述了所采用的基于专利实体和实体关系来完成对专利的内涵描述，并采用前后向网络构建的算法来计算每一个评价因子的定量分析，进而实现对专利的整体分析与评估。以目前国内石墨烯制备专利为例，应用所构建的专利创新指数进行了分析。基于专利实体和实体关系来完成对专利的内涵描述，并采用前后向网络构建和相似性分析的算法来计算每一个评价因子的定量分析。目前的研究是在当前静态数据库的基础上划分训练数据和测试数据，对于未来时间序列上的动态分析方法是技术创新管理领域的另一个重要问题。通过动态的专利分析技术，实现专利演进过程的实时分析与刻画，是接下来技术创新管理的重点发展方向。通过动态研究，能够更准确地反映技术创新的全过程。另外，本文的研究注重专利文本的挖掘使用，但是对多源大数据融合的技术路径分析尚未完成。未来，融合科技数据、商业数据、政策环境数据和企业报表数据等商业与市场数据的技术分析是一个重要研究方向。通过大数据的融合研究能够为技术创新提供更准确的分析维度。

本文工作得到重庆工商大学引进高层次人才科研启动项目 (1956053) 的支持，在此表示感谢。

参 考 文 献

[1] 余丰. 专利摘要的信息抽取研究[D]. 北京: 北京理工大学, 2006.
YU Feng. Study on information extraction of patent summary[D]. Beijing: Beijing Institute of Technology, 2006.

[2] 周俏丽, 蔡东风, 张桂平. 面向英文专利文本单语模板的自动抽取方法[J]. 沈阳航空工业学院学报, 2010, 27(4): 37-40.
ZHOU Qiao-li, CAI Dong-feng, ZHANG Gui-ping. Automatic acquisition approach of monolingual translation template oriented to English patent text[J]. Journal of Shenyang Institute Aeronautica Engineering, 2010, 27(4): 37-40.

[3] 姜彩红, 乔晓东, 朱礼军. 基于本体的专利摘要知识抽取[J]. 现代图书情报技术, 2009, 2: 23-28.
JIANG Cai-hong, QIAO Xiao-dong, ZHU Li-jun. Ontology-based patent abstracts' knowledge extraction[J]. New Technology of Library and Information Service, 2009, 2: 23-28.

[4] 吴志祥, 王昊, 王密平. 中文专利术语层次关系解析研究[J]. 情报学报, 2017, 4: 40-50.
WU Zhi-xiang, WANG Hao, WANG Mi-ping. A study on Chinese patent terms hierarchy parse[J]. Journal of the China Society for Scientific and Technical Information, 2017, 4: 40-50.

[5] 王密平, 王昊, 邓三鸿. 基于 CRFs 的冶金领域中文专利术语抽取研究[J]. 现代图书情报技术, 2016, 6: 28-36.
WANG Mi-ping, WANG Hao, DENG San-hong. Extracting Chinese metallurgy patent terms with conditional random fields[J]. New Technology of Library and Information Service, 2016, 6: 28-36.

[6] 孙凌云. 面向产品概念设计的专利地图技术研究[D]. 杭州: 浙江大学, 2008.
SUN Ling-yun. Research on patent mapping technology for product conceptual design[D]. Hangzhou: Zhejiang

University, 2008.

[7] GHOULA N, KHELIF K, DIENG K R. Supporting patent mining by using ontology-based semantic annotations[C]// IEEE AVIC/ACM International Conference on Web Intelligence. Fremont: IEEE, 2007: 131-139.

[8] TSENG Y, LIN C, LIN Y. Text mining techniques for patent analysis[J]. Information Processing & Management, 2007, 43(5): 1216-1247.

[9] JIN B, TENG H, SHI Y. Chinese patent mining based on sememe statistics and key-phrase extraction[J]. Advanced Data Mining and Applications, 2007, 46(32): 516-523.

[10] HAN H, ZHU L, ZHANG Z. Extracting hierarchical relationship of scientific and technical terms from unstructured text[J]. Natural Language Engineering, 2014, 25(6): 77-89.

[11] AJCRES L, YANG Y. Text mining and visualization tools[J]. World Patent Information, 2008(30): 280-293.

编 辑 叶 芳