计算机科学
COMPUTER SCIENCE

基于深度学习与文本计量的技术趋势分析

韦入铭，陈若愚，李晗，刘旭红

引用本文

韦入铭，陈若愚，李晗，刘旭红.基于深度学习与文本计量的技术趋势分析[J]．计算机科学, 2022, 49(11A): 211100119-6.

WEI Ru-ming, CHEN Ruo-yu, LI Han, LIU Xu-hong. Analysis of Technology Trends Based on Deep Learning and Text Measurement [J]. Computer Science, 2022, 49(11A): 211100119-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100119

基于深度学习与文本计量的技术趋势分析

韦入铭[1]　陈若愚[1]　李　晗[1]　刘旭红[1,2]
1 北京信息科技大学数据科学与情报分析研究所　北京 100101
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
(ruming_wei@163.com)

摘　要　传统的技术趋势分析工作需要由经验丰富的从业者完成,涉及到大量的文献调研和分析,工作耗时耗力。针对上述问题,提出一种基于深度学习与文本计量的技术趋势分析模型,设计基于BERT_BiLSTM_CRF模型的领域文献命名实体识别算法,优化BERT的掩码机制。以集成电路领域的新闻和论文为数据集,开展BiLSTM_CRF、BERT_BiGRU_CRF等模型以及文中所提BERT_BiLSTM_CRF* 模型的对比研究,研究命名实体识别技术在集成电路等领域的数据识别效果。相比于其他算法,文章所提的领域文献命名实体识别算法在F1值上达到了88.6%,奠定了技术趋势分析的基础。基于知识图谱易表达关联关系的特点,创新性提出知识图谱与文本计量技术结合的方法,并从不同角度以可视化的形式展示技术趋势分析效果,最终辅助从业者开展技术趋势智能分析工作。

关键词:命名实体识别;知识图谱;BERT_BiLSTM_CRF;文本计量;技术趋势分析
中图法分类号　TP391

Analysis of Technology Trends Based on Deep Learning and Text Measurement

WEI Ru-ming[1],CHEN Ruo-yu[1],LI Han[1] and LIU Xu-hong[1,2]
1 Laboratory of Data Science and Information Studies,Beijing Information Science and Technology University,Beijing 100101,China
2 Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,Beijing Information Science and Technology University,Beijing 100101,China

Abstract　Traditionally,technical trend analysis tasks need to be done by experienced analysts,involving a lot of literature review and data analysis work,which is time-consuming and labor-intensive. Facing the above problems,this paper proposes a technology trend analysis model based on deep learning and text measurement,and a domain specific named entity recognition(NER) algorithm based on the BERT_BiLSTM_CRF model is designed with optimized masking mechanism. Taking news and literatures texts in the field of integrated circuit as data set,a comparative study between BiLSTM_CRF,BERT_BiGRU_CRF and the optimized BERT_BiLSTM_CRF* model proposed in this paper is carried out. The performance of NER is compared and analyzed. Compared with other algorithms,the proposed algorithm reaches 88.6%(measured by F1 value),laying the foundation for technical trend analysis. Based on the characteristics of knowledge graphs that relationships can be naturally expressed,an innovative method that combines knowledge graphs with text measurement technology is proposed,and the results of technical trend analysis are visualized from various perspectives,and ultimately assist analysts to carry out intelligent analysis of technical trends.

Keywords　Named entity recognition,Knowledge graph,BERT_BiLSTM_CRF,Text measurement,Technology trend analysis

1　引言

领域技术趋势分析有助于从业者把握领域的发展方向,缩小行业发展差距。传统的技术趋势分析需要由经验丰富的从业者完成,涉及到大量的文献调研和分析,工作耗时耗力。因此,提出一种基于人工智能的领域技术趋势分析算法,以期减少人工投入,提高分析效果。

文章结合深度学习、文本计量、知识图谱等先进技术设计技术趋势分析模型。以BERT_BiLSTM_CRF领域文献命名实体识别算法为基础,通过优化传统的BERT掩码机制,运用

基金项目:北京信息科技大学勤信人才项目(2021);促进高校分类发展-重点研究培育项目——适应智慧城市应用场景的本体深度信念网络模型构建研究(2121YJPY225);科研机构创新能力建设-数据科学与情报分析研究所;促进高校内涵发展——面向边缘计算的创新科研平台建设项目(2020KYNH105)
This work was supported by the Qin Xin Talents Cultivation Program,Beijing Information Science & Technology University(2021),Promoting the Development of University Classification-key Research and Cultivation Projects—research on the Construction of an Ontology Deep Belief Network Model Suitable for Smart City Application Scenarios(2121YJPY225),Innovation Capacity Building of Scientific Research Institutions-Institute of Data Science and Information Analysis,Promote the Development of the Connotation of Colleges and Universities—an Innovative Scientific Research Platform Construction Project for Edge Computing(2020KYNH105).
通信作者:陈若愚(ruoyu-chen@foxmail.com)

BiLSTM 的门机制关注领域文献的实体信息，使用 feature function 抽象表达特征，从专利、论文、新闻、政策、研报 5 个维度出发，提取非结构化文本信息中的命名实体，实现领域文献命名实体的识别。相比于其他模型，本文所提的领域命名实体识别效果在 F1 值上分别有不同程度的提升。基于领域命名实体识别模型效果，结合知识图谱与文本计量等相关技术，举例分析集成电路、智能制造等多个领域的发展趋势，最终以可视化的形式展示分析效果。

文章提出的方法具有普适性，并且可以减少大量人工投入，辅助从业者理清特定领域的历史发展趋势以及当前发展的宏观态势，提升我国短板行业的国际地位和竞争态势。

2 国内外研究现状

技术趋势分析一直是各界研究热点。Wang[1] 基于病毒特征以及传统反病毒技术存在的缺陷，分析新一代反病毒技术趋势的发展；Gupta 等[2] 通过分析约翰霍普金斯大学的存储库中 30 个时间段的数据，采用时间序列预测方法预测新冠疫情未来在印度的发展趋势。Li[3] 以无线通信作为切入点，通过分析无线通信技术的热点及应用，研究通信技术的发展趋势。上述研究者从不同角度对不同领域进行技术趋势分析，但是存在 3 个共同的缺陷：1）研究者进行趋势分析时大多以常年积累的理论知识为基础，结合现有材料进行趋势分析，而经验不足的新从业者难以入手；2）研究者分析国内外某行业的发展趋势时并没有给出清晰明了的行业发展数据，难以支撑其提出的相关论点；3）研究者多从单一维度出发，如只关注某国对产业的影响，而很少关注时间的变化对行业的影响、行业顶尖公司对行业的影响、行业顶级研究者对行业的影响等，故上述论文中所提出的结论比较片面。

本文提出的技术趋势分析方法的重点在于使用命名实体识别技术提取专利、论文、新闻、政策、研报等非结构化文本中的命名实体，也是技术趋势分析方法的基础。

命名实体识别（Named Entity Recognition，NER），又称“实体识别”，是信息提取的子任务，旨在从非结构化的文本数据中识别出具有特定含义的实体。在 MUC-6[4] 之前，命名实体识别主要用于识别人名、地名、机构名三大类名词。从 MUC-6 开始，研究人员开始针对特定领域进行专有名词识别的研究，如医疗领域的命名实体识别[5]、军事领域的命名实体识别[6] 等。

命名实体识别是当前自然语言处理领域的研究热点。其发展从早期的基于词典和规则的方法[7]，到基于统计学习的方法[8]，再到目前比较受欢迎的混合方法。

基于词典和规则的学习方法通过设计相应的正则表达式或者规则，使用词典把符合模式的字符串匹配出来，作为命名实体识别任务的结果，但该方法可执行性差，且需要不同领域的语言学专家参与规则制定，容易产生歧义。后续出现了基于统计学习的方法。在基于统计学习的方法中，命名实体识别被当作序列标注问题。通过对大规模语料的学习得出标注模型，随后利用标注模型对句子中每个字或词进行标注。但该方法对语料库的依赖性较强，而现有的可用于各领域命名实体识别任务的通用语料库较少。由于基于词典和规则的学习方法和基于统计学习的方法在单独使用时都有一定的缺陷，所以研究学者开始重点研究利用混合方法来解决命名实体识别的问题。常用的混合方法主要有：1）统计学习方法之间或者内部层叠融合；2）机器学习与词典、规则融合；3）各类模型、算法相结合，即把前一级模型的输出当作后一级模型的输入，训练出新一级模型。由于在大多数情况下混合方法的有效性和准确性均优于传统的机器学习方法，所以该方法在目前的命名实体识别领域具有很高的地位。常见的混合模型主要有：BERT_BiLSTM_CRF，LATTICE_LSTM[9]，CNN_LSTM_CRF[10] 等。

其中，在命名实体识别任务中，BERT_BiLSTM_CRF 模型应用最为广泛，并且在很多领域都取得了不错的效果。如 Gao[11] 在一般数据集上构建了一个基于 BERT 的 BiLSTM_CRF 命名实体识别模型，并在 BiLSTM 和 CRF 之间增加了 Attention 网络，以协助对 CRF 的命名实体识别进行序列化；Guo[12] 构建了基于 BERT_BiLSTM_CRF 的法律案件实体智能识别算法，并通过实验验证了模型的有效性，提高了法律案件的处理效率。

但是 BERT_BiLSTM_CRF 是基于英语背景提出的，为了在保留英文文本上下文信息的同时不泄露标签信息，随机选择句子序列进行字级别的掩码处理，从而全面提取句子的特征。对于中文而言，词语蕴含的信息往往比字蕴含的信息更丰富、更准确，掩码机制以字级别处理中文文本会有明显的缺陷，所以本文对 BERT 预处理模型进行了优化，以词为单位做句子序列掩码处理，使其在处理中文文本时也具有很好的效果。

本文提出基于 BERT_BiLSTM_CRF*（优化 BERT 掩码机制后的模型，下文同）模型的领域文献命名实体识别算法，以集成电路、智能制造等领域的新闻与论文摘要为源数据，研究 BERT_BiLSTM_CRF* 模型在领域文献中的命名实体识别效果。实验结果表明，BERT_BiLSTM_CRF* 模型在测试集上的 F1 值为 85.29%，通过与 BERT_BiLSTM_CRF（原始 BERT 模型）模型[13] 和 BiLSTM_CRF 模型[14] 进行对比，结果表明本文所提模型的 F1 值最高。

将领域文献命名实体识别算法提取出来的领域名词、国家、时间等实体存放在知识图谱中。知识图谱可以利用图的形式表达客观世界中的概念、实体以及实体与概念之间的复杂关系，由谷歌[15] 最先提出，并被广泛应用于语义搜索、问答系统、智能决策、规则推理等智能领域。基于此，本文使用知识图谱存储领域实体之间的概念、关系及属性等信息是不错的选择。

3 技术趋势分析模型设计

3.1 模型总体设计

技术趋势分析模型由领域文献命名实体识别技术、TF-IDF、知识图谱、数据可视化等模块组成。如图 1 所示，文章以新闻及论文摘要作为训练数据，训练出具有领域特性的命名实体识别算法模型；然后使用该算法抽取专利、政策、研报中的领域实体，使用识别结果计算领域文献命名实体与文章中的 TF-IDF 值，取出文章最优关键词；再将最优关键词输入知识图谱，构建领域知识库；最后以可视化的形式展示出来。

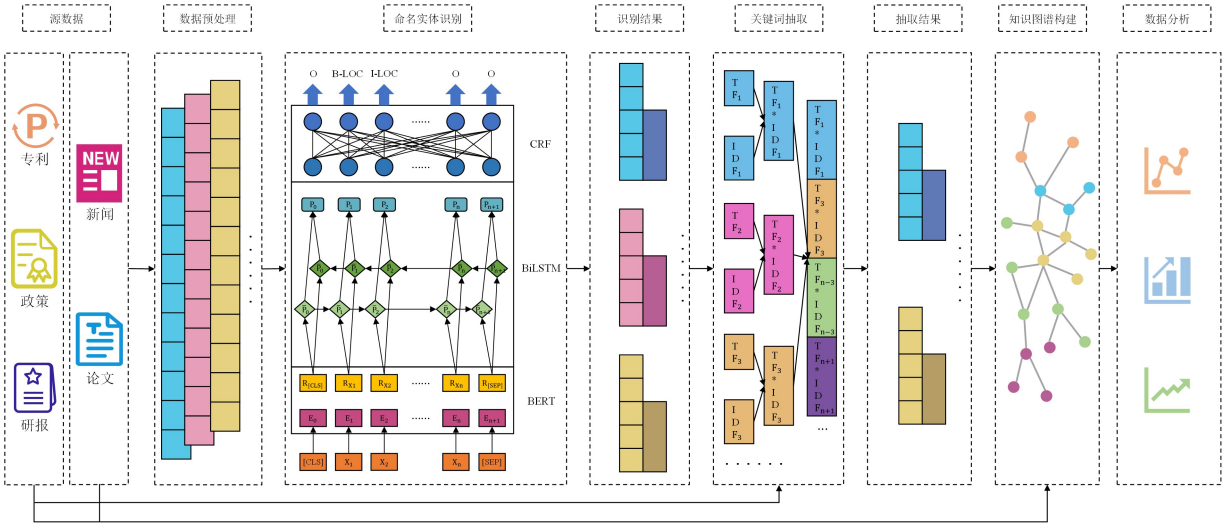

图 1　趋势分析模型的总体设计

Fig. 1　Overall design of trend analysis model

3.2　文本预处理

采用新闻及论文摘要作为命名实体识别模型的训练数据。由于源数据的长度不等，格式不统一，需要对存在残缺数据、错误数据、重复数据的源数据进行处理，将不符合标准的数据进行删除；部分数据有效信息稀疏，需要对稀疏数据进行去停用词等操作，增加实体识别模型的泛化能力，提高模型效率。

3.3　领域文献命名实体识别

命名实体识别和分类问题不完全相同，分类问题只需要判断出文本的词性/类型，具有一定的独立性。而命名实体识别对前后文的语义信息具有较强的依赖性。

论文提出的 BERT_BiLSTM_CRF* 模型由一个带有 Token 的 BERT 和用于解析字向量的 BiLSTM 模型以及一个线性链 CRF[16] 组成。BERT 采用 Transformer 模型作为编码器，通过词向量将输入的文本向量化作为模型输入，模型的输入除了需要的词向量外，还需要用于刻画全局语义信息的句向量和记录词在文本中所在位置的位置向量。通过优化 BERT 的掩码机制，使其更适用于中文本文序列。针对领域文献实体上下文联系较强的问题，使用 BiLSTM 层对输入序列做语义编码处理，得到预测标签序列。添加 CRF，利用转移矩阵保证输出标签之间的顺序性，从而提高预测的准确率。模型结构如图 2 所示。

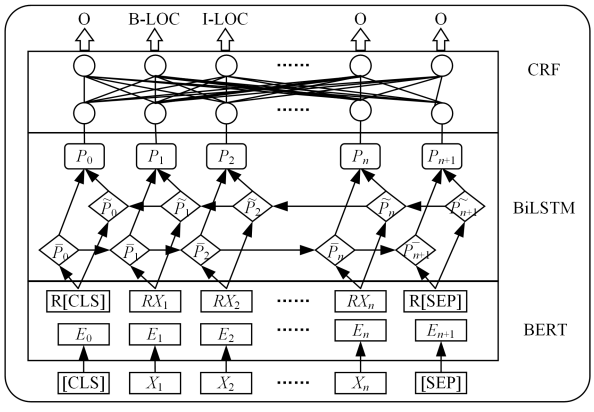

图 2　BERT_BiLSTM_CRF 原理图

Fig. 2　Schematic diagram of BERT_ BiLSTM_ CRF

3.3.1　BERT 层优化

在中文使用掩码机制时，首先对长文本进行分词处理，如果词语的某一部分被 Mask 掉，则设定该词的其他部分也同时会被 Mask，从而最大程度地保存了词语的语义信息。处理效果如表 1 所列。

表 1　优化后的掩码处理效果

Table 1　Optimized mask processing effect

Original text	智能制造行业的就业前景
Text segmentation	智能制造 行业 的 就业 前景
Original Mask	[Mask]能制造 行业 的 就业 前景
Optimize Mask	[Mask] [Mask] [Mask] [Mask] 行业 的 就业 前景

对于输入的文本序列，在序列的开头和结尾分别加入[CLS][SEP]标识符进行分割，然后再对句子进行优化 Mask 处理，得到带有位置信息、词信息以及句子信息的模型输入。

采用多头注意力机制（Multi_Head Attention）向量化输入语料。所谓多头，其实就是做了多次 attention 计算，如图 3 所示。

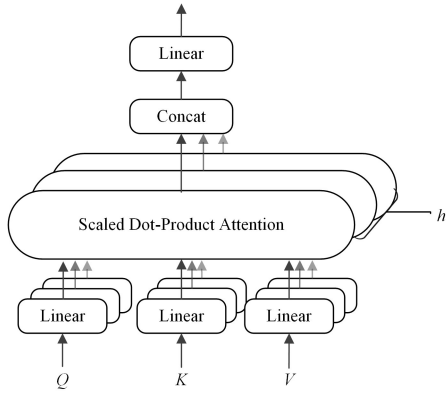

图 3　多头注意力机制原理图

Fig. 3　Schematic diagram of multi-head attention mechanism

Attention 主要用于获取领域文献句子序列中词与词的依赖关系，如式(1)所示：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 是指字向量矩阵，d_k 是向量矩阵维度。将多层 Attention 结果拼接起来可以得到不同空间各领域文献词语的位置信息及上下文的语义信息，如式(2)和式(3)所示：

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}) \tag{2}$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_n)\boldsymbol{W}^0 \quad (3)$$

其中，$\boldsymbol{W}$ 表示权重矩阵。

3.3.2 BiLSTM 层优化

长短期记忆网络(Long Short-Term Memory，LSTM)，可以较好地捕捉长距离句子的依赖关系，在训练长文本时，可以智能地通过门机制保留有效信息而抛弃无用信息，增强模型的泛化能力。专利、论文、新闻、政策、研报等文本均为长文本，可以使用 LSTM 捕捉领域文献的依赖关系。LSTM 门机制结构如图 4 所示。

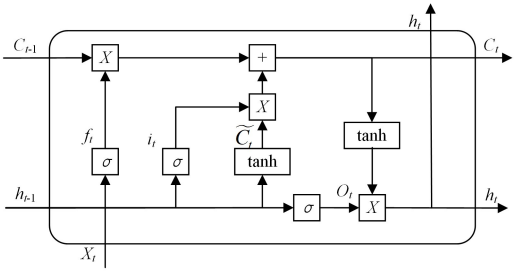

图 4 BiLSTM 门机制原理图

Fig. 4 Schematic diagram of BiLSTM gate mechanism

将前一时刻的隐藏状态 h_{t-1} 与当前时刻的领域文献名词 X_t 输入遗忘门 f_t，抛弃无用信息，如式(4)所示：

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \quad (4)$$

将前一时刻的隐藏状态 h_{t-1} 与当前时刻的领域文献名词 X_t 输入记忆门 i_t，保留需要记忆的重要信息，并输出细胞的临时状态 $\widetilde{C}_t$，如式(5)－式(6)所示：

$$i_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_i) \quad (5)$$

$$\widetilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \quad (6)$$

获取 f_t，i_t，$\widetilde{C}_t$ 之后，结合上一时刻的细胞状态 C_{t-1} 计算当前时刻的细胞状态 C_t，如式(7)所示：

$$C_t=f_tC_{t-1}+(1-f_t)\widetilde{C}_t \quad (7)$$

最终计算输出门及当前时刻的隐层状态，得到与句子长度一致的隐层状态序列 $\{h_0,h_1,\cdots,h_{n-1}\}$。输出门及当前时刻的隐层状态如式(8)、式(9)所示：

$$o_t=\sigma(\boldsymbol{W}_o\cdot[h_{t-1},x_t]+\boldsymbol{b}_o) \quad (8)$$

$$h_t=o_t\tanh(C_t) \quad (9)$$

其中，σ 表示激活函数，$\boldsymbol{W}$ 表示权重矩阵，$\boldsymbol{b}$ 为偏置向量。

LSTM 只能按照单一方向学习语义知识，而很多情况下，要想完全理解句子的含义，需要结合上下文信息，所以产生了 BiLSTM 模型。BiLSTM 就是把前向 LSTM 和后向 LSTM 相结合，如通过前向 LSTM 获取“智能制造”“的”“就业”“前景”的前向向量序列 $\{\boldsymbol{h}_{L0},\boldsymbol{h}_{L1},\boldsymbol{h}_{L2},\boldsymbol{h}_{L3}\}$，通过后向 LSTM 获取“前景”“就业”“的”“智能制造”的后向向量序列 $\{\boldsymbol{h}_{R0},\boldsymbol{h}_{R1},\boldsymbol{h}_{R2},\boldsymbol{h}_{R3}\}$，最后将前向和后向向量序列进行拼接得到 $\{[\boldsymbol{h}_{L0},\boldsymbol{h}_{R3}],[\boldsymbol{h}_{L1},\boldsymbol{h}_{R2}],[\boldsymbol{h}_{L2},\boldsymbol{h}_{R1}],[\boldsymbol{h}_{L3},\boldsymbol{h}_{R0}]\}$，即 $\{\boldsymbol{h}_0,\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{h}_3\}$。

3.3.3 CRF 层优化

BiLSTM 输出标签需要考虑上下文的语义信息，对前后文的语义信息具有较强的依赖性，如“B-PER”(人名开始)后面一定是“I-PER”(人名中间)，不可能是其他标签。CRF 可以获取标签间的依赖关系，输出最优预测序列，弥补 BiLSTM 存在的不足。

对于输入序列 $X=(x_1,x_2,\cdots,x_n)$ 和预测标签序列 $Y=(y_1,y_2,\cdots,y_n)$，得出该标注序列的得分为：

$$S(X,Y)=\sum_{i=0}^{n}A_{y_iy_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \quad (10)$$

其中，$P_{i,j}$ 表示第 i 个字符被预测为标签 j 的概率，$\boldsymbol{A}$ 表示序列的转移矩阵，$A_{y_iy_{i+1}}$ 表示序列标签从 y_i 转移到 y_{i+1} 的得分。该模型训练出正确标签的最大对数概率为：

$$\log(p(Y|X))=s(X,Y)-\log\sum\widetilde{Y}\in Y_X e^{s(X,\widetilde{Y})} \quad (11)$$

其中，Y_X 表示所有可能的标签序列。式(11)中的结果由动态规划计算所得，在评估中通过 Viterbi 解码得到最可能的标注序列。

3.4 技术趋势分析算法

技术趋势分析算法的重点在于找到文章的核心词，挖掘文章与核心词、核心词与核心词以及文章与文章的关联关系。

论文提出的技术趋势分析算法是基于领域文献命名实体开展的，由 TF-IDF 和知识图谱组成。TF-IDF 挖掘文章的核心词，知识图谱用于建立词与文章的关系。

TF-IDF 由 TF(Term Frequency，词频)及 IDF(Inverse Document Frequency，逆向文件频率)组成，旨在提取文章的主题词，评估词语对于文本的重要程度。TF 是指某个词语在一篇文章中出现的次数。文本语料库中出现某个词的文档个数越大，IDF 值越小。IDF 计算公式如式(12)所示：

$$IDF=\log\left(\frac{\text{语料库中文档的总数}}{\text{包含某个词的文档数}+1}\right) \quad (12)$$

其中，分母之所以加 1，是为了避免分母为 0 的情况。

采集到文章主题词之后，将文章主题词与文献命名实体识别算法中提取到的国家、作者、机构等重要信息输入到知识图谱中，构建领域知识库。

4 实验设置与结果分析

4.1 数据采集与预处理

相对于结构化数据，非结构化文本数据包含的有效信息密度较低，但权威媒体新闻、论文资源等公开数据质量较高，可以避免价值密度低的缺陷，且相对容易获取，数据样本大，而且涉及诸如财政、国防、金融、医疗、通信、邮电、交通等较为广泛的领域。

文本实验数据来源于中国新闻网、新华网、科技日报等主流新闻网站的 4000 余篇新闻，以及中国知网、万方数据库的 1000 余篇论文摘要。

实验中共划分 4 类实体：人名(PER)、地名(LOC)、机构名(ORG)以及领域文献命名实体(PRO)。以 BIO(B：Begin，I：Inside，O：Outside)标注体系对数据集进行标注，领域文献命名实体的划分以“全国科学技术名词审定委员会”审定正式公布的为标准。标签设定如表 2 所列。

表 2 标签含义

Table 2 Label meaning

Label	Meaning
B-PER	人名开头
I-PER	人名中间
B-LOC	地点开头
I-LOC	地点中间
B-ORG	组织开头
I-ORG	组织中间
B-PRO	领域文献命名实体开头
I-PRO	领域文献命名实体中间
O	其他

数据处理后的格式每一行由一个“字”、一个“空格”以及一个“标签”组成，且定义一行为一条字级别标注数据。

经过数据清洗等预处理操作后，共有约 300 万条字级别标注数据，其中约 294 万条数据为训练集，约 3 万条数据为

验证集，约 3 万条数据为测试集（分别占总数据的 98%，1%，1%）。为了保证数据的有效性，在机器标注的基础上对 3 万条测试集数据进行人工校对。

除 BERT_BiLSTM_CRF* 以外，本实验还比较了在同一数据集下 BERT 和 BiLSTM 等主流模型的效果，并采用 word2vec[17] 训练模型所需要的词向量。

4.2 领域文献命名实体识别结果分析

实验通过精确率、召回率、F1 值评估 BERT_BiLSTM_CRF* 与 BERT，BiLSTM，BiLSTM_CRF，BERT_IDCNN_CRF，BERT_BiGRU_CRF 以及 BERT_BiLSTM_CRF 的命名实体效果。为了减小实验误差，文章实验所取结果均为 3 次实验中 F1 值的平均值。

实验 1　比较 BERT_BiLSTM_CRF* 与目前主流模型的命名实体识别结果。本实验将在自行采集的数据中进行评估。

由表 3 可知，BERT_BiLSTM_CRF* 模型在准确率、召回率上均优于其他传统的机器学习模型。通过分析实验结果，得出可能原因如下：BERT_BiLSTM_CRF 在命名实体识别的 F1 值上比 BiLSTM_CRF 高 3.03%，说明 BERT 预处理阶段对实验最终的结果影响较大，有助于提高命名实体识别的 F1 值；BERT_BiLSTM_CRF 相比于 BERT_IDCNN_CRF 以及 BERT_BiGRU_CRF 两种主流模型，其 F1 值分别提高了 1.52%与 0.9%，说明在本文数据集，BERT_BiLSTM_CRF 更有优势，这也是本文选择 BERT_BiLSTM_CRF 作为元模型的最主要原因；本文优化模型 BERT_BiLSTM_CRF* 在 F1 值上比 BERT_BiLSTM_CRF 提高了 0.76%，说明在中文数据集上，BERT 层优化有效。

表 3　不同模型的命名实体识别效果

Table 3　Named entity recognition effect of different models

（单位：%）

Model	*Precision*	*Recall*	*F1*
BiLSTM	79.93	80.41	80.17
BERT	83.24	83.89	83.56
BiLSTM_CRF	84.70	84.92	84.81
BERT_IDCNN_CRF	87.52	85.15	86.32
BERT_BiGRU_CRF	88.16	85.75	86.94
BERT_BiLSTM_CRF	88.35	87.33	87.84
BERT_BiLSTM_CRF*	**89.67**	**87.55**	**88.60**

实验 2　比较 BERT_BiLSTM_CRF* 与目前主流模型在相同数据集下的领域专有名词识别效果。

在实验 1 中，BERT_BiLSTM_CRF* 的效果全面优于传统的机器学习模型，所以此实验侧重于比较相同数据集下，BERT_BiLSTM_CRF* 与传统模型在专有名词命名实体间的识别效果。实验结果如表 4 所列。从表 4 中可以看出，BiLSTM 添加了 CRF 层后在本文数据集上的效果有了明显的提升，说明 CRF 的序列校正工作起到了至关重要的作用，并一举超过了单一的 BERT 模型；值得注意的是，优化后 BERT_BiLSTM_CRF* 有着明显的优势，并且相比于优化前的 BERT_BiLSTM_CRF 模型在 F1 值上也提高了 3.46%。通过分析实验结果与实验数据，专有名词多为 4～8 个字符长度的词语，优化后的 BERT 层在保留词语完整性的同时保留了词语的语义信息，使得最终的识别效果有所提高，优化后的 BERT 更适合处理中文领域文献命名实体，所以导致 BERT_BiLSTM_CRF* 领域文献命名实体识别效果在本文数据集下优于其他模型效果。

表 4　模型在专有名词上的识别效果

Table 4　Recognition effect of model on proper nouns

（单位：%）

Entity	Model	*Precition*	*Recall*	*F1*
PRO	BiLSTM	72.64	73.70	73.17
	BERT	77.89	80.56	79.20
	BiLSTM_CRF	80.27	85.44	82.77
	BERT_IDCNN_CRF	85.65	83.04	84.32
	BERT_BiGRU_CRF	84.07	85.12	84.59
	BERT_BiLSTM_CRF	86.12	85.39	85.75
	BERT_BiLSTM_CRF*	90.08	88.35	**89.21**

4.3 技术趋势分析效果呈现

效果 1　基础领域知识库构建效果展示

以关键词“半导体”为中心节点展开，可以看到半导体关联的领域为“智能制造”，同时展示了知识库中和半导体相关的所有信息，包括 92 篇论文、1 个机构和 25 个相关政策等信息，如图 5 所示。

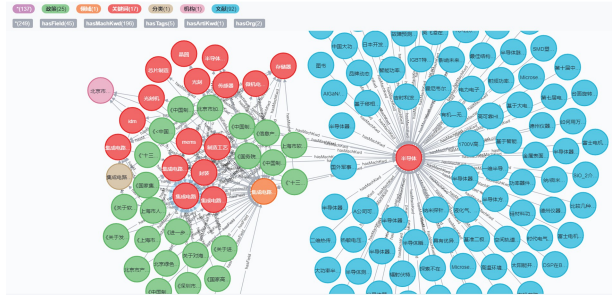

图 5　知识库效果图

Fig. 5　Knowledge graph rendering

效果 2　集成电路领域技术发展趋势分析

以 2016－2020 年集成电路领域为例，筛选出专利、论文、新闻、政策、研报 5 个维度的所有相关非结构化文本数据。通过图 6，可以很清晰地看出从 2016－2020 年集成电路制造领域的发展变化，可以有助于研究者清晰明了地把握研究热点，及时了解国内外集成电路制造领域的发展现状。

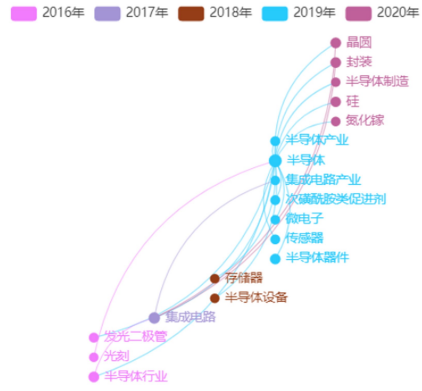

图 6　集成电路制造领域的技术发展趋势图

Fig. 6　Technology development trend chart in the field of integrated circuit manufacturing

效果 3　智能机器人领域企业发展趋势分析

以智能服务机器人领域为例，通过对国际各大公司论文的发表数量以及相关新闻的报道数量进行计量，绘制出国际公司在 2015－2020 年 TOP10 的折线图，如图 7 所示。对折

线图的趋势进行分析，可以推测出某公司对智能服务机器人的关注力度与技术发展情况，推动国家把握国际公司的发展趋势，辅助决策人员进行智能决策。

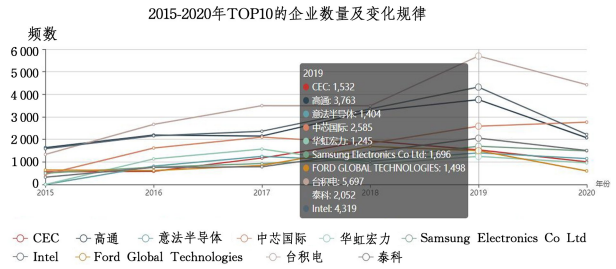

图 7 智能机器人领域的企业发展趋势图

Fig. 7 Development trend of enterprises in the field of intelligent robots

结束语 本文实现了 BERT_BiLSTM_CRF* 在领域文献的命名实体识别算法。与 BERT_BiLSTM_CRF 以及 BiLSTM_CRF 等其他模型相比，BERT_BiLSTM_CRF* 在本实验环境下的领域文献命名实体的识别效果更优。

领域文献的命名实体识别研究有助于快速构建领域知识图谱，整合散落在网络空间里的领域材料，建立领域数据间的互联互通，运用文本计量方法把握技术趋势。

领域文献命名实体识别算法效果的好坏将直接影响领域文献分析的效果，因此在今后的研究中，将继续推进领域文献命名实体识别的研究，提升识别效果。

参 考 文 献

[1] WANG C Y. Analysis on the development trend of anti-virus technology [N]. Network World, 2005-12-12(035).

[2] GUPTA R, PAL S K. Trend Analysis and Forecasting of COVID-19 outbreak in India[J]. MedRxiv, 2020.

[3] LI S W. Analysis of the hot spots and development trend of communication technology in the era of big data[J]. Information Recording Materials, 2021, 22(7): 62-64.

[4] GRISHMAN R, SUNDHEIM B. Message Understanding Conference 6: A Brief History[C]// Proceedings of the 16th International Conference on Computational Linguistics. 1996.

[5] ZHAO S, LIU T, ZHAO S, et al. A neural multi-task learning framework to jointly model medical named entity recognition and normalization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 817-824.

[6] XUEZHEN Y I N, HUI Z, JUNBAO Z, et al. Multi-neural network collaboration for Chinese military named entity recognition [J]. Journal of Tsinghua University (Science and Technology), 2020, 60(8): 648-655.

[7] XIE R, LIU Z, JIA J, et al. Representation Learning of Knowledge Graphs with Entity Descriptions [C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[8] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence data[C]// Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). 2001: 282-289.

[9] ZHAO S, CAI Z, CHEN H, et al. Adversarial training based lattice LSTM for Chinese clinical named entity recognition [J]. Journal of Biomedical Informatics, 2019, 99: 103290.

[10] WU F, LIU J, WU C, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C]// The World Wide Web Conference. 2019: 3342-3348.

[11] GAO X, LI Q. Named entity recognition in material field based on Bert-BILSTM-Attention-CRF [C] // 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS). IEEE, 2021: 955-958.

[12] GUO Z X, DENG X L. The entity intelligent identification method of legal cases based on BERT-BiLSTM-CRF [J]. Journal of Beijing University of Posts and Telecommunications, 2021, 44(4): 129-134.

[13] GU Y. Research on Complex Chinese Named Entity Recognition Based onBiLSTM-CRF [D]. Nanjing: Nanjing University, 2019.

[14] HU H, DENG S, LU H, et al. A Comparative Study on the Classification Performance of Machine Learning Models for Academic Full Texts [C] // International Conference on Information. Cham: Springer, 2020: 713-737.

[15] TIAN LX. Summary of Research on Knowledge Graph[J]. Software, 2020, 41(4): 67-71.

[16] LIU S H, LIU X H, LIU X L, et al. Extraction of coal mine safety accident ontology concept based on word vector and conditional random field[J]. Coal Technology, 2018, 37(9): 178-181.

[17] GOLDBERG Y, LEVY O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv: 1402. 3722, 2014.

WEI Ru-ming, born in 1997, postgraduate. His main research interests include natural language processing and knowledge graph.

CHEN Ruo-yu, born in 1982, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include natural language processing, data mining, semantic network and so on.