

基于 BERT 嵌入 BiLSTM-CRF 模型的 中文专业术语抽取研究

吴俊¹, 程垚¹, 郝瀚¹, 艾力亚尔·艾则孜², 刘菲雪¹, 苏亦坡¹

(1. 北京邮电大学经济管理学院, 北京 100876; 2. 深圳暴风智能科技有限公司, 北京 100191)

摘要 专业术语的识别与自动抽取对于提升专业信息检索精度, 构建领域知识图谱发挥着重要基础性作用。为进一步提升中文专业术语识别的精确率和召回率, 提出一种端到端的不依赖人工特征选择和领域知识, 基于谷歌BERT预训练语言模型及中文预训练字嵌入向量, 融合BiLSTM和CRF的中文专业术语抽取模型。以自建的1278条深度学习语料数据为实验对象, 该模型对术语提取的F1值为92.96%, 相对于传统的浅层机器学习模型(如左右熵与互信息算法、word2vec相似词算法等)和BiLSTM-CRF深度神经网络模型的性能有较为显著的提升。本文也给出了模型应用的具体流程, 能够为中文专业术语库的构建提供实践指南。

关键词 BERT; BiLSTM; CRF; 专业术语抽取

Automatic Extraction of Chinese Terminology Based on BERT Embedding and BiLSTM-CRF Model

Wu Jun¹, Cheng Yao¹, Hao Han¹, Ailiyaer·Aizezi², Liu Feixue¹ and Su Yipo¹

(1. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876;

2. Shenzhen Storm Intelligent Technology Co., Ltd, Beijing 100191)

Abstract: High quality professional term recognition and its extraction play an important role in the fields of domain information retrieval and knowledge graph building. To improve the precision and recall rate of terminology recognition, we propose a Chinese terminology recognition and extraction approach that does not rely on specific domain knowledge or artificial features. Using the latest developments in representation learning, this study introduces BERT embedding as a character-based pre-trained model and incorporates it with a bi-directional long short-term memory (BiLSTM) and a conditional random field (CRF) to extract deep learning terminologies from 1278 annotated datasets collected from domain e-books. The experimental results show that the proposed model reaches 92.96% in F-score and outperforms other competing algorithms, such as left and right entropy, mutual information, a word2vec based similar terminology recognition algorithm, and a BiLSTM-CRF model. The best practices and related procedures for the implementation of the proposed model are also provided to guide its users in its further improvement.

Key words: BERT; BiLSTM; CRF; terminology recognition and extraction

收稿日期: 2019-10-10; 修回日期: 2019-10-30

基金项目: 国家重点研发计划项目“基于模式创新的科技咨询服务研发与应用示范”(2018YFB1403600); 北京市社会科学基金一般项目“基于大数据的北京市共享单车产业监测与发展趋势研究”(17YJB018)。

作者简介: 吴俊, 男, 1971年生, 副教授, 硕士生导师, 研究方向为文本挖掘与服务创新; 程垚, 女, 1996年生, 硕士研究生, 研究方向为数据分析与服务创新, E-mail: chengyao@bupt.edu.cn; 郝瀚, 男, 1998年生, 本科生; 艾力亚尔·艾则孜, 男, 1991年生; 刘菲雪, 女, 1999年生, 本科生; 苏亦坡, 女, 1999年生, 本科生。

1 引言

专业术语的识别、抽取与术语库的构建是中文自然语言处理 (natural language processing, NLP) 领域的重要问题, 对于提升专业信息检索精度、构建领域知识图谱发挥着重要基础性作用。一般认为, 领域术语的抽取可以转化为序列标注 (sequence labelling) 问题, 可以应用命名实体识别 (named entity recognition, NER) 的研究思路与方法来解决。在 NER 领域, 既有的研究方法大多从两条路径出发: 一是基于文本的统计分布特征, 从多词 (字) 共现视角, 引入左右熵与互信息算法或 word2vec 词嵌入等, 通过术语的上下文词语共现特征完成术语的识别; 另一种是基于深度神经网络针对训练集数据开展基于词语级或字符级的术语领域语义特征学习, 进而实现对测试集数据的术语识别与抽取。近年来, 深度神经网络在新闻等通用语料领域的命名实体识别表现出了较好的性能。相比于传统基于规则或统计机器学习的方法, 基于神经网络的深度学习方法更加具有泛化能力强、更少依赖人工特征选择的优点。尤其是基于双向长短时记忆网络 (bi-directional long short-term memory, BiLSTM) 和条件随机场 (conditional random field, CRF) 算法成为提升命名实体识别效果的重要途径。

Huang 等^[1]基于拼写特征构建了 BiLSTM-CRF 模型, 在 CONLL2003 英文语料上取得了 F1 值 88.83% 的效果。陈伟等^[2]基于 BiLSTM-CRF 模型对爱奇艺线上图文视频短标题进行关键词自动抽取, 在简单关键词层取得了 86.7% 的 F1 值。陈世梅等^[3]在中文否定与不确定信息语料上, 应用 BiLSTM-CRF 模型进行触发词识别和覆盖域识别, 分别取得了 91.03% 和 73.91% 的 F1 值, 且实验结果表明相对于 CRF 模型和 BiLSTM 模型, BiLSTM-CRF 模型在汉语否定触发词识别和覆盖域识别上效果更优。

需要指出的是, BiLSTM-CRF 模型在通用命名实体识别领域取得了不错的效果, 但在专业领域命名实体识别时仍然存在以下一些难点: ①专业领域的中文术语实体存在多种表述方式, 大都没有统一的命名方法或标准术语库; ②待识别的术语实体可能由多个字符或词语构成, 实体边界难以划分; ③待识别的术语实体存在缩写、嵌套、中英文混合等情况。因此, 专业领域的中文术语实体识别仍旧依赖人工界定的特征和领域专业知识, 使得术语的识别精确率和召回率受领域情境的限制而无法推广应用。

近年来针对语言模型的研究表明, 丰富的、无监督的预训练是语言理解系统不可或缺的组成部分, 能够有效改进包括自然语言推断、复述 (paraphrasing) 等句子层面的任务, 以及命名实体识别、SQuAD 问答等“词” (token) 层面的任务^[4,6]。2013 年 Google 提出的 word2vec 词嵌入方法将“词” (token) 映射为单一向量, 尽管可以表征一义多词, 但无法解决一词多义问题。ELMo (embeddings from language models) 模型^[5]在 2018 年由 Matthew E. Peters 等人提出, 该论文获评当年 NAACL (NLP 领域顶级会议之一) 最佳论文, 相比于 word2vec, 较好地解决了表征“词” (token) 的多义性问题。进一步地, 在 ELMo 模型基础上, 2018 年 10 月, Google 公司的 Devlin 等提出 BERT (bidirectional encoder representations from transformers) 模型^[6], 采用表征语义能力更强的双向 Transformer 编码器来对海量公开语料预训练, 由此形成的预训练字向量在用于下游的众多英文 NLP 任务中表现出超强的性能。

鉴于 BERT 预训练语言模型在英文 NLP 任务中的优异表现, 本文尝试在中文命名实体识别任务中引入 BERT 中文预训练向量, 在此基础上, 提出一种端到端的不依赖领域知识和人工特征的 BERT-BiLSTM-CRF 模型, 该模型首先利用 BERT 中文预训练向量将语料文本转为字符级嵌入向量, 然后将其送入 BiLSTM 深层神经网络进行训练, 之后利用 CRF 对 BiLSTM 的输出进行处理, 获得全局最优的术语标记序列。与已有研究相比, 一方面, 本文提出的以 BERT 字符嵌入向量为基础, 融合 BiLSTM 和 CRF 的术语识别与抽取模型, 相对于浅层机器学习模型 (如左右熵与互信息算法、word2vec 相似词算法等) 和 BiLSTM-CRF 深度神经网络模型有显著的性能提升, 在测试数据集上 F1 值达到 92.96%; 另一方面, 本文给出了从专业领域电子文档中抽取语料, 应用开源文本标注工具提升人工标注效率, 借助人工智能云服务商在线 GPU 算力完成模型构建与训练的实施路径, 也为广大研究者和企业实践者应用本文提出的方法提供了实践指南。

2 专业术语提取相关研究

从专业文本语料中识别并抽取领域术语是一项富有挑战性的工作, 主要体现在: 一方面, 专业书籍中的领域术语专业性很强, 大多未建立国家标准, 缺乏统一的规范, 较少出现在通用词库中; 另一方面, 专业书籍中术语缩略语、实体包含、互指

现象也较为普遍,这对识别术语的正确性、完整性提出了更高要求。本节首先对传统基于文本统计特征的术语提取算法及其应用进行回顾,指出存在的优点与缺点,之后简要说明基于深度神经网络的深度学习算法与应用,进一步指出本文研究方法的主要改进之处。

2.1 基于统计特征的术语提取算法与应用

多词(字)表达(multiword expressions, MWEs)是指2个或2个以上的词(字)单元共现概率相对较高、结合紧密且非合成性比较强的一类词(字)组合^[7],代表着若干词(字)之间的紧密结合关系。从统计学的视角看,多词表达内部词语之间的结合紧密程度依赖于词语之间的共现频率,也就是说多词表达的一个字串往往以较高的概率与另外一部分共同出现。因此,基于统计特征的专业术语提取可以视为多词(字)表达的识别。在这方面,左右熵和互信息算法最具代表性。左右熵(left and right entropy, LRE)能够体现词(字)串结合的不确定程度,而互信息(mutual information, MI)主要体现两个词之间的相互依赖程度。融合左右熵和互信息算法来识别并提取专业术语已被用于上市公司年报风险短语的识别^[8]、互联网文本的多词表达抽取^[9]等。这种方法的优点是算法完全基于字词的统计特征,实施简单且高效;不足之处在于术语的识别精度较低,召回率不高。

从统计语言模型发展而来的分布式表示词嵌入(word embedding)方法以word2vec最为受人关注^[10-11]。word2vec将one-hot表示的词语稀疏向量转化为低维度的连续稠密向量,可以在词向量中包含更丰富的上下文语义信息。基于训练得到的词向量,可以计算两两词语之间的关系,如词语相似性、语义关联性等。因此,将左右熵和互信息算法识别的专业术语作为种子词,通过应用word2vec,也可找到语义相近的同类专业术语。但该方法也存在如下缺陷:每一个种子词及其相似词语彼此之间通常呈现较高的余弦相似度,使得获取的候选术语词串收敛在一定范围内,不够全面。

2.2 基于深度神经网络的深度学习算法与应用

在专业术语识别过程中,传统的浅层机器学习方法大多需要人工制定规则或选择一组适合该任务的特征模板,算法识别的好坏依赖于规则的合理性或特征模板的质量,较为费时费力。深度神经网络(deep neural network, DNN),尤其是具有时序学习

特点的循环神经网络(recurrent neural network, RNN)、长短时记忆网络(long short-term memory, LSTM)等,采用基于词向量的特征表示,把词向量作为深度神经网络的输入,能够自动学习文本上下文深层语义信息,并识别术语边界,大大减少了对人工特征和领域知识的依赖程度,有效地促进了NER性能的提升。

李丽双等^[12]用CNN-BiLSTM-CRF模型,通过CNN卷积获得字符级的词形态特征,与词嵌入向量组合,在Biocreative II GM和JNLPBA2004生物医学专业语料上分别达到了89.09%和74.40%的高F1值。李健龙等^[13]应用CNN-BiLSTM-CRF模型解决军事领域的命名实体识别问题,通过CNN模型对字向量矩阵进行卷积和池化处理,将字向量拼接到词向量上后使用获得的字词结合向量训练BiLSTM模型,最后利用添加了注意力机制的BiLSTM模型对军事领域命名实体进行识别,获得87.38%的F1值。李明浩等^[14]针对中医医案临床症状术语,提出了一种基于LSTM和CRF的深度学习症状术语识别方法,在基于LSTM-CRF序列标注模型的基础上,针对中医症状特征添加了部分字符级别的要素特征,在以古汉语为主的小规模中医医案语料上最高取得了78%的F1值。冯艳红等^[15]基于4种中文文本语料,对人名、地名、机构名进行命名实体识别,将传统的机器学习方法CRF和SVM(support vector machine)、基于深度神经网络的DNN模型、BiLSTM模型4种方法进行对比,并证实依次加入词向量E、标签约束和领域知识后BiLSTM模型识别效果逐步提升,且超过传统的命名实体识别方法中的最优结果。张应成等^[16]应用包含词向量层、BiLSTM网络层、神经网络语言模型CRF层3层结构的BiLSTM-CRF模型,以50000条招标平台上的招标文件为语料,对招标人、招标编号、招标代理进行了识别,最好的F1值达到87.86%。他的研究也进一步指出,BiLSTM方法优于LSTM方法,且引入CRF算法可以给不同模型带来程度不等的效果提升。

综上所述,在基于深度神经网络的模型中,BiLSTM-CRF模型逐渐成为命名实体识别领域公认较好的方法,不同的学者针对各自关注的专业领域语料,或者通过词向量嵌入,或者增加领域特征,均取得了接近90%的F1值性能。不过这些方法也存在着以下不足:①仍然依赖领域专业词库或专业知识特征,模型的跨行业、跨领域泛化应用能力不强;②对于识别精确率和召回率要求较高的场景(如专业知识库的构建),还有进一步提升空间。

3 基于BERT字嵌入的BiLSTM-CRF术语抽取模型

随着NLP领域深度神经网络模型研究的深入，不依赖人工特征的端到端解决办法日益成为主流。本文以近年NER领域较为主流的BiLSTM-CRF模型作为术语识别的基准模型，通过构建基于BERT字嵌入向量的BiLSTM-CRF模型来抽取中文专业书籍中的术语，并比较它与基准模型以及传统的左右熵+互信息术语识别算法、基于word2vec的相似词术语识别算法的性能。之所以选择引入BERT字嵌入预训练向量作为神经网络的输入，主要基于以下考虑：由于中文存在字和词的不同粒度划分，相应也存在基于字的NER、基于词的NER和字词结合的NER这3种方案，已有研究表明基于字符的NER一般有更好的表现^[4,17]，而BERT采用的是基于字符的预训练方案，在英文NER任务中表现突出，因此可以进一步检验其对中文NER任务的适用性。

3.1 模型整体框架

BERT-BiLSTM-CRF模型整体结构如图1所示，整个模型分为3个部分，先通过BERT预训练向量获得输入语料字符的语义表示，得到句子中每个字

的向量表示之后，再将字向量序列输入BiLSTM中进一步语义编码，最后通过CRF层输出概率最大标签序列。

与其他基于深度学习的命名实体识别模型相比，该模型最主要的区别是加入了BERT预训练中文向量，BERT预训练向量由Google在中文大规模语料上学习所得，可以通过文本上下文计算字符的向量表示，能够表征字符的多义性，增强句子的语义表示。该模型在实际使用时固定BERT预训练向量参数，只训练BiLSTM-CRF参数，这种训练方式可以相对减少训练参数，缩短训练时间。

3.2 BERT预训练语言模型

近年来的不少研究指出，学习广泛适用的词嵌入表示是现代NLP系统的一个不可分割的部分，经过预训练的词嵌入要比传统的one-hot编码导入有利于后续任务性能的改进。ELMo模型^[5]从不同的维度对传统的词嵌入进行了概括，通过构建词嵌入动态调整的双向神经网络，能够提取上下文敏感特征，输出体现上下文语义的预训练词向量，较好地解决了一词多义问题。BERT^[6]相比于ELMo模型进一步拓宽了词向量的泛化能力，能够充分学习字符级、词级、句子级甚至句间关系特征，增强字向量

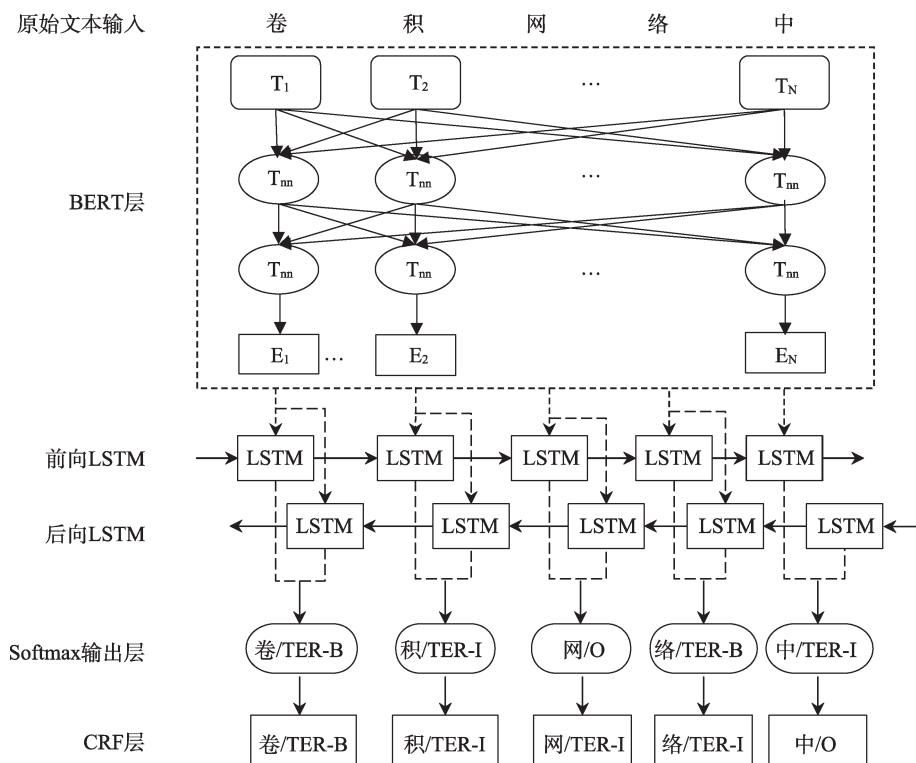


图1 BERT-BiLSTM-CRF模型示意图

的语义表示，因此表现出优于过往方法的卓越性能^[4,17]。

传统的RNN和CNN在处理NLP任务时存在着一定缺陷：RNN的循环式网络结构没有并行化，训练慢；CNN先天的卷积操作不是很适合序列化的文本。Transformer模型^[18]是文本序列网络的一种新架构，基于self-attention机制，任意单元都会交互，没有长度限制问题，能够更好捕获长距离上下文语义特征。BERT模型采用了多层的双向Transformer编码器结构，同时受到左右语境的制约，相比于ELMo模型中独立训练从左到右和从右到左的LSTM连接，能够更好地包含丰富的上下文语义信息。此外，Transformer针对self-attention机制无法抽取时序特征的问题，采用了位置嵌入的方式来添加时序信息，BERT输入表示由标记嵌入（词嵌入）、句子嵌入和位置嵌入这3个嵌入拼接，能够在一个标记序列中清楚地表示单个文本句子或一对文本句子，如图2所示。

此外，BERT预训练语言模型还通过masked language model（遮蔽语言模型）和next sentence prediction（下一句预测）两个任务分别捕捉词级和句子级的表示，并进行联合训练。遮蔽语言模型是为了训练深度双向语言表示向量的模型，通过随机遮蔽句子里某些单词，然后预测被遮蔽的单词，类似于“完形填空”的学习模式。相比于传统标准语言模型只能从左至右或从右至左单向预测目标函数，遮蔽语言模型可以从任意方向预测被掩盖的单词。

下一句预测则是为了训练一个理解句子关系的模型，由于许多重要的NLP下游任务，如问题回答（question answer, QA）和自然语言推理（natural language interaction, NLI），都是建立在理解两个文本句子之间的关系的基础上，而语言模型不能很好地直接产生这种理解，因此，该任务通过预训练一个二分类的模型（随机替换一些句子，再基于上一句进行IsNext/NotNext的预测），来学习句子之间的

关系。

BERT预训练模型与其他语言模型相比，具有更强的上下文长距离语义学习能力，相应生成的字嵌入分布式表示可以用于下游NLP任务，尤其是专业术语识别这一类NER任务中。

3.3 BiLSTM及CRF

在序列标注任务中，循环神经网络（RNN）由于能够保存文本序列的历史信息，成为解决序列标注问题的首选。相比于RNN模型，长短时记忆网络（LSTM）^[19]在隐藏层加入了特别设计的记忆单元，能够较好解决RNN在训练中由于序列过长而产生的梯度弥散和梯度消失问题，从而应用于命名实体识别任务中。一个典型的LSTM网络结构可以形式化表示如下：

$$i_t = \sigma(x_t \cdot w_{xh}^i + h_{t-1} \cdot w_{hh}^i + b_h^i) \quad (1)$$

$$f_t = \sigma(x_t \cdot w_{xh}^f + h_{t-1} \cdot w_{hh}^f + b_h^f) \quad (2)$$

$$o_t = \sigma(x_t \cdot w_{xh}^o + h_{t-1} \cdot w_{hh}^o + b_h^o) \quad (3)$$

$$\tilde{c}_t = \tanh(x_t \cdot w_{xh}^c + h_{t-1} \cdot w_{hh}^c + b_h^c) \quad (4)$$

$$c_t = i_t \otimes \tilde{c}_t + f_t \otimes c_{t-1} \quad (5)$$

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

式中， σ 是激活函数sigmod； \otimes 表示点乘运算； \tanh 表示双曲正切激活函数； x_t 是单元输入； i_t 、 f_t 、 o_t 分别表示在 t 时刻的输入门、遗忘门和输出门； w 、 b 分别代表输入门、遗忘门和输出门的权重矩阵和偏置向量； \tilde{c}_t 表示 t 时刻的状态，是仅由当前输入得到的中间状态，用于更新当前时刻的状态； h_t 表示 t 时刻的输出。

与LSTM相比，双向长短时记忆网络（BiLSTM）对每个句子分别采用顺序（从第一个词开始，从左往右递归）和逆序（从最后一个词开始，从右向左递归）计算得到两套不同的隐层表示，然后通过向量拼接得到最终的隐层表示。因此，BiLSTM能够更好地捕捉双向的语义依赖关系，有效地学习并掌握上下文语义共现信息，从而提升命名实

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token embeddings	E _[CLS]	E _{my}	E _{dog}	E _{is}	E _{cute}	E _[SEP]	E _{he}	E _{likes}	E _{play}	E _{##ing}	E _[SEP]
Segment embeddings	+ E _A	+ E _B									
Position embeddings	E ₀	E ₁	E ₂	E ₃	E ₄	E ₅	E ₆	E ₇	E ₈	E ₉	E ₁₀

图2 BERT输入向量表示

体识别的性能。

尽管 BiLSTM 能够有效地识别实体的边界，但有时不考虑标记的实体序列之间关系。为此，可以引入机器学习中的条件随机场 (CRF) 算法，来获得全局最优的标记序列。CRF 是以给定的随机变量为输入，求解输出随机变量条件概率分布的算法，近年来被频繁应用于词性标注、句法分析和命名实体识别领域^[12]。BiLSTM 不考虑标签之间的相关性，而 CRF 的一个独特优势是能够通过考虑相邻标签的关系获得一个全局最优的标记序列，基本算法如下。

对于一个句子 $S = \{W_1, W_2, \dots, W_n\}$ 送入网络中训练，定义矩阵 P 是 BiLSTM 层的输出，其中 P 的大小为 $n \times m$ ， n 是单词个数， m 是标签类别。定义 P_{ij} 代表句子中第 i 个单词的第 j 个标签的概率。对于一个预测序列 $y = \{y_1, y_2, \dots, y_n\}$ ，它的概率可以表示为

$$K(X, y) = \sum_{i=0}^n A_{y_i y_{i+1}} + \sum_{i=1}^n P_{iy_i} \quad (7)$$

式中，矩阵 A 是转移矩阵； A_{ij} 表示由标签 i 转移到标签 j 的概率； y_0, y_n 是预测句子起始和结束的标记。 A 是一个大小为 $m+2$ 的方阵，在原始句 S 的条件下产生标记序列 y 的概率为

$$p(y|S) = \frac{e^{K(X,y)}}{\sum_{\tilde{y} \in Y_X} K(X, \tilde{y})} \quad (8)$$

式中， \tilde{y} 代表真实的标记值。在训练过程中标记序列的似然函数为

$$\log(p(y|S)) = K(X, y) - \log(\sum_{\tilde{y} \in Y_X} e^{K(X, \tilde{y})}) \quad (9)$$

式中， Y_X 表示所有可能的标记集合，包括不符合 BIO (beginning-inside-outside) 三元标注规则的标记序列。通过式(9)得到有效合理的输出序列。预测时，由式(10)输出整体概率最大的一组序列：

$$y^* = \arg \max_{\tilde{y} \in Y_X} K(X, \tilde{y}) \quad (10)$$

因此，将 CRF 与 BiLSTM 神经网络相结合，对 BiLSTM 的输出进行处理，可以获得最佳的术语标记结果。

4 实验过程及结果分析

4.1 语料的人工标注

为检验提出方法的实践应用价值，本文选取专业领域的中文语料进行分析。研究选取《深度学习

500 问》^①电子书作为深度学习领域专业术语构建的文本语料对象。选取原因如下：首先，近年来深度学习在语音、图像、自然语言处理等领域取得了不错的成果，是人工智能领域的前沿热点，该书以问答形式对常用的概率知识、线性代数、机器学习、深度学习、计算机视觉等热点问题进行阐述，具有较好的专业广度与深度，能够反映出深度学习领域的最新术语动态；其次，该电子书开源共享，方便下载和文本数据的处理。

语料的采集步骤如下：首先，抽取该电子书的第 1~3 章共 48374 字文本，作为人工标注语料对象。其次，经研究组人员共同讨论，设置针对此语料的术语注释规则：①该术语和机器学习、深度学习领域相关。在人工标注前，先让标注员学习机器之心整理的人工智能术语表^②，对领域的常见术语有一定了解与熟悉，如果不是与机器学习、深度学习领域密切相关，即便是专业词语，也不在标注范围内。②该词语应该是专业术语/词语，只有在机器学习、深度学习领域内才常遇见，其他场景较少遇见。

语料的人工标注采取严格的标准流程：

Step1. 标注员粗略阅读一遍分发的文本语料，对语料背景有初步印象。

Step2. 标注员浏览一遍机器之心的人工智能术语表，熟悉机器学习和深度学习领域的一些常见术语。

Step3. 标注员背靠背独立完成分发语料专业词语/术语的标注。

Step4. 根据标注结果，计算标注员两两之间的一致性程度（用卡帕系数衡量）。

Step5. 标注员两两核对各自标注原稿，协商标注产生差异的原因，如果可以达成共识，则形成一致的标注结果，如果不能，对标注歧义的文本用颜色标注。

Step6. 指导教师对标注员标注结果进行复核，确认标注结果形成最终的标注语料。

为确保上述标注步骤的规范化，提高标注工作效率，研究人员在腾讯云服务器上部署了开源文本注释工具软件 doccano。该软件可以实现文本命名实体注释、文本分类注释和序列到序列 (sequence to sequence, seq2seq) 任务注释等常用文本标注功能。从深度学习电子书籍中标注深度学习术语属于文本命名实体注释范畴，在 doccano 中应用主要包括四步骤，分别是项目构建、导入数据、术语标注和导出数据，如图 3 所示。

^①访问 <https://github.com/scutan90/DeepLearning-500-questions> 可以获取。

^②访问 <https://github.com/jiqizhixin/Artificial-Intelligence-Terminology> 可以获取。



图3 应用dooccano标注深度学习术语步骤示意

整个数据集的注释由两名标注人员在老师指导下按照上述注释规则与流程进行，注释结果 κ 值为0.87，可以认为两人的注释结果高度一致。将导出的已标注术语数据集按照7:3比例划分为训练集和验证集数据，语料数据集统计信息如表1所示。

表1 数据集统计信息

分类	句子数	字符数	术语实体数
训练集	882	68979	3319
验证集	396	30251	1432

在此基础上，采用 BIO 三元标注法对数据集上的标注进行调整，其中，B (beginning) 表示当前字符是深度学习术语实体的首字，I (inside) 表示当前字符是深度学习术语实体的非首字，O (outside) 表示当前字符不是深度学习术语实体。表2给出了深度学习术语实体的示例标注。

表2 深度学习术语示例标注

而	且	弱	分	类	器
O	O	B-TERM	I-TERM	I-TERM	I-TERM
构	造	极	其	简	单
O	O	O	O	O	O

4.2 数据处理过程

为比较 BERT-BiLSTM-CRF 方法与其他多种术语识别与抽取算法的性能，研究人员将数据处理分为两阶段展开。

阶段一，针对基于统计特征的术语提取，采用了基于互信息和左右熵的术语提取算法以及基于 word2vec 词语相似度的术语提取算法。

基于互信息+左右熵的术语提取算法实现步骤如下：首先，对预处理完毕并且通过 jieba 分词后的语料应用 N-Gram 模型，分别取 $N=1,2,3$ ，将所有得到的词串组及对应出现概率存入字典模型中。然后，遍历字典所有的二阶共现词组，根据概率计算得到其互信息值，通过阈值筛选掉一部分二阶共现，将合格的共现词加入候选术语集中。最后，针对这些候选术语，进一步计算其左右熵，并叠加互信息和左右熵的值得到每一个二阶共现词的分值，取分值降序 TopN 术语，与人工标注的术语进行比对。

基于 word2vec 的词语相似度术语提取算法实现步骤如下：首先，对预处理完毕并且通过 jieba 分词后的语料，统计并计算每一个词语的 tf-idf 值，根据降序筛选出 Top20 的关键词作为算法种子词。然后，取窗体大小为 5，向量维度为 100，对语料应用 word2vec 词向量模型，并利用模型的最相似方法计算词向量空间中，与每一个种子词余弦相似度最大的 100 个子词语，将得到的种子词与子词语加入候选术语集。遍历候选集中词语的二元组合，也加入候选术语集中。最后，对比人工标注的术语并统计候选术语集的命中情况。

阶段二，针对基于深度学习的术语提取算法，本文的实施过程采用典型的机器学习流程。将数据

集按照 7 : 3 比例划分为训练集和验证集，针对训练集进行模型训练，基于验证集检验模型性能。本阶段采用的实验环境、实验过程和参数设置详述如下。

1) 实验环境

本文采用的实验环境如表 3 所示。

表 3 实验环境

操作系统	
CPU	Openbayes ^① 算力容器 2CPU 模式
GPU	NVIDIA T4(16 GB)
Python	3.6
TensorFlow	1.12.0
Keras	2.2.4
内存	30 G

2) 实验过程

BERT-BiLSTM-CRF 模型的训练采用固定 BERT 参数，只更新 BiLSTM-CRF 参数的特征提取方法。为了证明模型的有效性，与以下模型进行了对比实验：①BiLSTM-CRF 模型，该模型是序列标注领域表现较好的模型，对输入字符序列上下文语义进行学习，通过 CRF 输出全局最优的标记序列；②word2vec-BiLSTM-CRF 模型，该模型使用 word2vec 代替 BERT 完成词向量训练工作，然后将 word2vec 输出的文本词嵌入向量导入 BiLSTM-CRF 神经网络训练。

3) 参数设置

Google 开源可供下载的 BERT 中文预训练向量参数如下：BERT-Chinese 一共 12 层，隐层为 768 维，采用 12 头模式，共 110 M 个参数。训练时，最大序列长度采用 512，train_batch_size 为 64，learning_rate 和 drop_out_rate 均为默认值，BiLSTM 隐藏层维数为 256。

4.3 实验结果及分析

为检验不同算法的性能优劣，本文采用精确率、召回率和 F1 值作为度量指标。

精确率（precision, P ）、召回率（recall, R ）、F1 值（F1-score, F ）指标定义为

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$F = \frac{2 * P * R}{P + R} = \frac{2 * TP}{2 * TP + FP + FN} \quad (13)$$

式中， P 表示正确预测为正类占所有预测为正类的总数的比例； R 表示正确预测为正类占样本中真正正类的总数的比例； F 为 P 和 R 的调和平均值，表达了二者的综合效果。本文即采取以上 3 个指标来评价正类的识别效果。

从表 4 可以看到，BERT+BiLSTM+CRF 模型的精确率、召回率和 F1 值表现最佳，不仅远远超出传统基于统计特征的术语提取算法，也强于基于深度神经网络的 BiLSTM+CRF 和 word2vec+BiLSTM+CRF 算法，说明 BERT 预训练语言模型生成的字向量能更好地表示字符的上下文语义信息。另外，在模型训练调参过程中，我们也发现，sequence length 在从 128、256 到 512 的调参过程中， P 、 R 和 F1 值都有一定的提升，说明序列长度参数设置对最终的识别性能有一定影响。

表 4 术语提取不同算法模型的性能比较

算法名称	精确率 (precision)	召回率 (recall)	F1 值
基于统计特征的术语提取算法			
1 左右熵+互信息	0.4110	0.3612	0.3845
2 基于 word2vec 的相似词	0.2312	0.2454	0.2381
1+2 的结合	0.6185	0.6564	0.6369
基于深度神经网络的术语提取算法			
1 BiLSTM+CRF	0.8623	0.8916	0.8767
2 word2vec+BiLSTM+CRF	0.6669	0.6033	0.6335
3 BERT+BiLSTM+CRF	0.9196	0.9342	0.9296

注：深度神经网络训练的 epochs 均设为 80，sequence length 设为 512。

表 4 实验结果中值得注意的是，针对字符型 BERT 预训练向量和词语型 word2vec 预训练向量加载 BiLSTM+CRF 的实验结果表明，加入词嵌入信息降低了模型的性能，但采用 BERT 字符嵌入向量能明显提升模型性能。这与 Meng 等^[20]发表于 ACL2019 的最新论文结论相印证，即在深度神经网络框架下开展中文 NLP 任务时，“字符”级别预训练向量的表现优于“词语”级别预训练向量的表现，可能的原因包括：①在用 word2vec 预训练词向量之前，采用 jieba 分词工具分词时，存在着专业术语切词不当、术语之间边界切分不准确的问题，导致术语不在词库（out-of-vocabulary, OOV）现象的发生；②经统计，在 1278 条总数据集中，共有词语

^① Openbayes 是位于北京的一家机器智能创业公司，提供基于 GPU 的算力容器服务。

54450个,训练集中,经jieba分词后的词语,出现一次的术语占训练集词数或者总词数的比例都在3%以内,说明训练数据集中的术语较为稀疏,由于词语数的增加,会使得神经网络模型参数增多,从而引起模型过拟合问题,造成模型在验证集数据上表现欠佳。

5 总 结

为提升领域专业术语的识别精确率,提高算法模型端到端部署的效率,本文提出了一种不依赖于领域知识和人工特征的BERT-BiLSTM-CRF方法,用于识别并提取专业术语。BERT预训练语言模型通过双向Transformer结构动态生成字符的上下文语义表示,比传统的词嵌入向量表示更能表征字符的语义与语句特征。通过采集的深度学习电子书籍语料的实验检验,该模型不仅大大优于基于文本统计特征的左右熵-互信息算法以及基于word2vec相似词算法,同时也优于NER领域表现较好的BiLSTM-

CRF模型。尤其需要指出的是,在本文的实验流程中,只需经过1278句样句的注释就取得了术语识别F1值92.96%的性能,不需要在模型中添加人工特征,因此具有较好的跨领域、跨行业应用前景。

本文的主要贡献或创新之处主要体现在三方面:首先,提出了一种不依赖人工特征选择和领域知识,基于谷歌BERT中文预训练向量,融合双向长短时记忆网络(BiLSTM)和条件随机场(CRF)的中文专业术语识别与提取算法模型。以自建的深度学习语料数据为对象,该模型对专业术语的提取精确度和召回率均优于传统基于统计文本特征和现有的BiLSTM+CRF算法模型,展现出较好的应用前景;其次,论文给出了从专业领域电子文档中抽取语料,应用开源文本标注工具提升人工标注效率,借助人工智能云服务商在线GPU算力完成模型构建与训练的实施路径,为经济管理领域的众多应用场景(如从商业文本中自动提取领域术语,从上市公司公告中自动提取风险术语等)提供了可复制的实践指南,如图4所示。

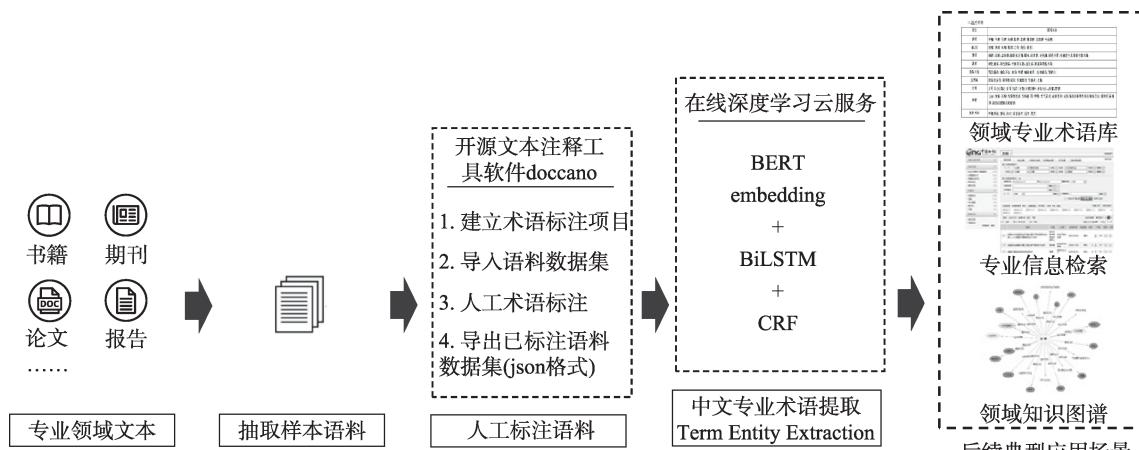


图4 应用深度学习模型自动提取专业术语路径与后续应用场景

此外,论文针对字符型BERT预训练向量和词语型word2vec预训练向量加载BiLSTM+CRF的实验结果表明,加入词嵌入信息降低了模型的性能,但采用BERT字符嵌入向量能明显提升模型性能。与Meng等^[20]发表于ACL2019的最新论文结论相印证,即在深度神经网络框架下开展中文NLP任务时,“字符”级别预训练向量的表现优于“词语”级别预训练向量的表现。

未来的研究一方面可以扩展其他领域的专业文本,以检验本文提出的方法及其优异性能在多个领域专业术语识别上的普适性;另一方面,也可以开

展BERT预训练字符向量用于其他神经网络性能的比较,例如,与BERT-CNN-LSTM-CRF、BERT-BiGRU-CRF等模型性能的比较等,进一步揭示不同模型的最佳应用场景与最优参数设置。

参 考 文 献

- [1] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[OL]. <https://arxiv.org/abs/1508.01991v1>.
- [2] 陈伟, 吴友政, 陈文亮, 等. 基于BiLSTM-CRF的关键词自动抽取[J]. 计算机科学, 2018, 45(S1): 91-96, 113.
- [3] 陈世梅, 伍星, 唐凡. 基于BiLSTM-CRF模型的汉语否定信息识别[J]. 中文信息学报, 2018, 32(11): 55-61.

- [4] 林怀逸, 刘箴, 柴玉梅, 等. 基于词向量预训练的不平衡文本情绪分类[J]. 中文信息学报, 2019, 33(5): 132-142.
- [5] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 2227-2237.
- [6] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.
- [7] Baldwin T, Bannard C, Tanaka T, et al. An empirical model of multiword expression decomposability[C]// Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. Stroudsburg: Association for Computational Linguistics, 2003: 89-96.
- [8] 胡小荣, 姚长青, 高影繁. 基于风险短语自动抽取的上市公司风险识别方法及可视化研究[J]. 情报学报, 2017, 36(7): 663-668.
- [9] 龚双双, 陈钰枫, 徐金安, 等. 基于网络文本的汉语多词表达抽取方法[J]. 山东大学学报(理学版), 2018, 53(9): 40-48.
- [10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[OL]. <https://arxiv.org/abs/1301.3781>.
- [11] Zhu L, Wang G J, Zou X C. Improved information gain feature selection method for Chinese text classification based on word embedding[C]// Proceedings of the 6th International Conference on Software and Computer Applications. New York: ACM Press, 2017: 72-76.
- [12] 李丽双, 郭元凯. 基于CNN-BLSTM-CRF模型的生物医学命名实体识别[J]. 中文信息学报, 2018, 32(1): 116-122.
- [13] 李健龙, 王盼卿, 韩琪羽. 基于双向 LSTM 的军事命名实体识别 [J]. 计算机工程与科学, 2019, 41(4): 713-718.
- [14] 李明浩, 刘忠, 姚远哲. 基于LSTM-CRF的中医医案症状术语识别[J]. 计算机应用, 2018, 38(S2): 42-46.
- [15] 冯艳红, 于红, 孙庚, 等. 基于BLSTM的命名实体识别方法[J]. 计算机科学, 2018, 45(2): 261-268.
- [16] 张应成, 杨洋, 蒋瑞, 等. 基于BiLSTM-CRF的商情实体识别模型[J]. 计算机工程, 2019, 45(5): 308-314.
- [17] 杨飘, 董文永. 基于BERT嵌入的中文命名实体识别方法[J/OL]. 计算机工程 (2019-05-30) [2019-07-10]. <https://doi.org/10.19678/j.issn.1000-3428.0054272>.
- [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Breach, 2017: 6000-6010.
- [19] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [20] Meng Y X, Li X Y, Sun X F, et al. Is word segmentation necessary for deep learning of Chinese representations? [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 3242-3252.

(责任编辑 魏瑞斌)