情报理论与实践
Information Studies:Theory & Application
ISSN 1000-7490,CN 11-1762/G3

《情报理论与实践》网络首发论文

题目：基于深度学习的领域本体概念自动获取方法研究
作者：王思丽，祝忠明，刘巍，杨恒
收稿日期：2019-09-09
网络首发日期：2019-10-28
引用格式：王思丽，祝忠明，刘巍，杨恒．基于深度学习的领域本体概念自动获取方法研究．情报理论与实践.
http://kns.cnki.net/kcms/detail/11.1762.G3.20191028.1429.006.html

网络首发时间：2019-10-28 15:17:44
网络首发地址：http://kns.cnki.net/kcms/detail/11.1762.G3.20191028.1429.006.html

●王思丽[1,2,3]，祝忠明[1,2,3]，刘　巍[1,2]，杨　恒[1,2]
（1．中国科学院西北生态环境资源研究院文献情报中心，甘肃 兰州 730000；2．中国科学院兰州文献情报中心，甘肃 兰州 730000；3．中国科学院大学，北京 100049）

基于深度学习的领域本体概念自动获取方法研究*

摘　要：[**目的/意义**]实现对领域概念的自动学习抽取，解决领域本体自动化构建的首要基础任务。[**方法/过程**]以无监督的学习方法和端到端的识别模式为理论技术基础，首先通过对主流词嵌入模型进行对比分析，设计提出了基于 Word2Vec 和 Skip-Gram 的领域文本特征词嵌入模型的自动生成方法；其次研究构建了以 IOB 格式的标注文本作为输入，基于自注意力机制的 BLSTM-CRF 领域概念自动抽取模型；最后以资源环境学科领域为例进行了实验研究与评估分析。[**结果/结论**]模型能够实现对领域概念的自动抽取，对领域新概念或术语的自动识别也具有一定的健壮性。[**局限**]模型精度尚未达到峰值，有待进一步优化提升。
关键词：深度学习；领域本体；概念自动获取；词嵌入；自注意力

Method of Domain Ontology Concept Automatic Extraction Based on Deep Learning

Abstract: [Purpose/significance] Realize the automatic learning extraction of domain concepts and solve the primary basic tasks of domain ontology automation construction.**[Method/process]** The unsupervised learning method and the end-to-end recognition mode are the theoretical and technical foundations. Firstly, through the comparative analysis of the mainstream word embedding model, the paper designs an automatic generation method of domain text feature word embedding model based on Word2Vec and Skip-Gram. Secondly, the paper constructs a domain concept automatic extraction model named BLSTM-CRF based on self-attention mechanism, using annotated text in IOB format as input. Finally, the paper takes the field of resources and environment as an example to carry out experimental research and evaluation analysis. [Result/conclusion] The model can realize the automatic extraction of the domain concepts, and it also has certain robustness to the automatic identification of new domain concepts or terms. [Limitations] The accuracy of the model has not yet reached the top value and needs to be further optimized.
Keywords: deep learning; domain ontology; concept automatic extraction; word embedding; self-attention

1 研究背景与意义

当前知识内容变革的步伐和范围进一步加剧，知识载体的多渠道、多来源、多格式、富媒体化、关联数据化[1]等复杂异构现象已成为常态，知识内在的分布性、非结构性、异质性已严重阻碍了知识在多主体间和软件实体间的共享和重用。领域本体作为一种基于语义来描述知识系统的概念模型的重要工具，具有强大的对概念特征的定义能力及对概念关系的描述能力，已成为知识挖掘和语义分析计算不可或缺的基础，被认为是大数据环境下解决“信息和知识孤岛问题”[2]的最佳方法。领域本体概念是指领域本体中能够代表某专业/主题领域知识的重要的相关的术语集合。从本体论的角度来看，领域概念是构成领域本体的必不可少的基础要素，是领域本体标引领域知识要点和表征领域知识分布及共性特征的重要体现。因而，实现领域本体概念的自动获取是领域本体自动构建研究的首要任务和基础工作。

传统的领域本体概念获取方法有基于（语言学）模式规则的方法、基于统计分析的方法、多种策略混合的方法等，但都存在相应不足。基于模式规则的方法

*本文为中国科学院兰州文献情报中心 2018 年主任基金项目“基于深度学习的领域本体自动构建方法研究”（项目编号：Y8AJ012005）和中国科学院 2019 年西部之光项目“开放学术资源的情景化组织与服务研究” (项目编号：Y9AX011001)的成果。

[3-6]常依赖人工进行浅层语法分析或领域词典构建模式规则进行概念识别与抽取，对特定语言的词典、标注语料库、语法库等先决资源条件依赖性大，存在规则维护/更新/扩展困难，应用范围有限，可移植性差等问题，尤其对一些新兴词汇、非正式句子、缩写短语、词典中没有的专业术语等识别较差，准确率和召回率低，因而无法大规模地应用于概念识别。基于统计分析的方法[7-10]一般是对大量领域文本数据进行统计分析，将满足统计阈值或条件的字符串序列作为领域概念，常用的统计方法有词频统计、TF-IDF、信息熵、互信息计算等，但该方法存在计算量大，常遗漏低频词，常忽略或缺乏上下文语义分析等问题，因而识别的准确率一直有待提高。为了突破上述局限，随着机器学习和自然语言处理技术的推动，后来大多数研究开始将领域本体概念获取问题规范化表述为概念抽取（Concept Extraction, CE）或术语抽取（Term Extraction, TE），并归入为一种序列标记的命名实体识别（Named Entity Recognition, NER）问题，主要采用的就是各种半监督和监督方式混合的机器学习算法及其变体等，聚焦于从领域文本中半自动或自动地获得领域依赖的属性、专门的文本特征、上下文语义信息等以解决上述问题。常用的混合策略方法[11-14]有如将层叠条件随机场（CRF）、支持向量机（SVM）算法及几种模式匹配规则结合起来构建的混合模型等；如将基于独热编码词特征表示方法的布朗聚类技术和隐马尔可夫模型（HMM）结合起来实现对未标记语料库的无监督特征表示模型[15-16]等；如将分布式词特征表示方法和随机索引模型结合起来，借助或直接将 Wikipedia、WordNet、HowNet、百度百科等的部分文本语料训练生成词嵌入模型以改善概念提取性能[17-18]。这些混合的策略方法常能够最大限度地减少依赖于词汇的查找的计算量，并逐步开始能够部分考虑利用上下文语义信息进行领域文本分析，一定程度上提高了识别的准确率和召回率。但传统机器学习方法本质上仍是遵循领域特定的特征工程和分类两个步骤，仍属于高度专用的手工制作系统范畴且需要劳动密集型的专家知识才能实施，需要大量的“经验”（专家知识）和“运气”（人工选取并获得最优特征的过程随机、难以复制且不可控）作为基础，因而难以大规模流行应用起来，自动精准识别领域本体中的概念仍是一项极具挑战意义的研究任务。

2 研究目标与内容

近年来，尤其是 2012 年以来，深度学习的出现和深度神经网络的激增，在计算机视觉、图像/语音识别等任务中已取得了前所未有的成果。但在 NER 及其相关扩展任务如领域本体的构建研究中，由于很多领域缺乏大量规范文本、标注语料、基础本体或词库等领域基础资源条件，深度学习的研究应用目前仍主要集中在生物医学领域和通用领域，其他领域还非常少。深度学习的主要优点是能够使用现成的或衍生的各种深度神经网络词嵌入模型或算法从领域文本中自动学习特征，从而避免了繁重且耗时的特征工程，且学习特征的过程是人工、领域、语言非依赖性的，是充分利用嵌入了上下文语义信息的，因而可移植、可重用、可扩展性也强。为了有效解决现有领域本体概念抽取方法中对人工定义特征的过度依赖和对隐含复杂特征的难以充分自动挖掘等问题，本文的主要研究目标是将领域本体概念获取问题转化为基于深度学习的自然语言处理中的序列标注概念识别问题，提出以 Word2Vec 词嵌入模型和 Skip-gram 算法为基础训练生成的无监督词嵌入作为初始化特征输入，通过联合改进经典深度学习算法长短期记忆网络 LSTM 和条件随机场算法（CRF）构建的基于自注意力机制的 BLSTM-CRF 领域本体概念自动抽取算法模型，并进行实验验证与性能评估。主要研究内容框架见图 1。

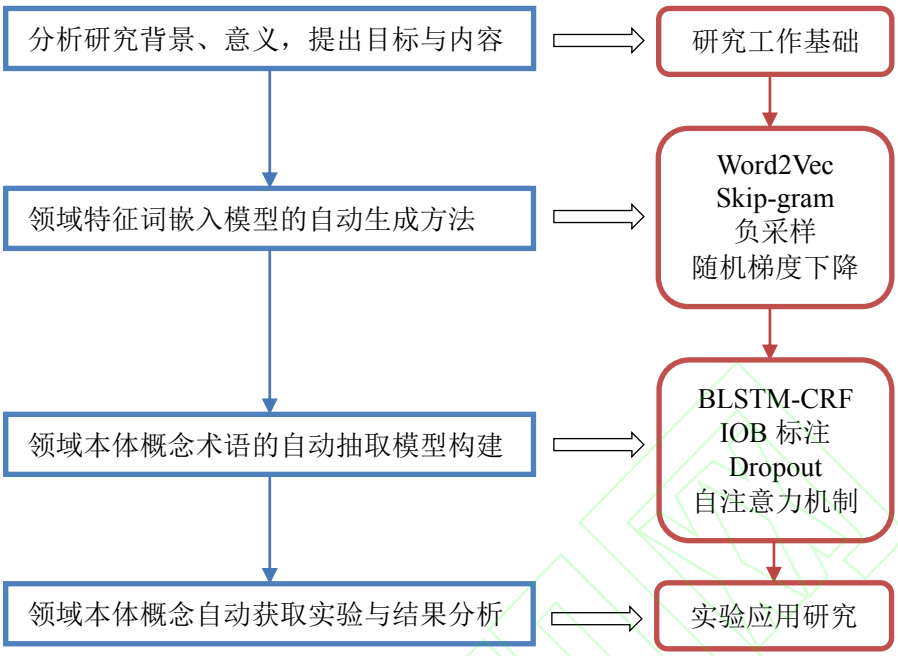

图 1 基于深度学习的领域本体概念自动获取研究框架

3 领域文本特征词嵌入模型的自动生成研究

3.1 几种主流词嵌入模型的对比分析

词嵌入是指利用自然语言处理的方法将文本中的单词转换和表示为连续的密集的向量，并能够保留和体现文本中词汇之间的语义和句法的相似性等特征。目前，词嵌入是深度学习最主要也几乎是唯一的初始化预输入数据形式，词嵌入模型的预训练生成已成为深度学习相关研究必须解决的首要任务。

实际上，以 2013 年谷歌首次公开发布 Word2Vec[19]模型引起广泛关注作为开端，近年来已有大量不同的词嵌入算法模型相继被开源提供，如 GloVe[20]，ELMo[21]，BERT[22]，XLNet[23]等。本文首先对上述几种主流词嵌入模型进行对比分析，见表 1。

表 1 几种主流词嵌入模型对比分析

模型（年份）	上下文	预训练方法	预测目标	下游任务	采样算法	向量级别
Word2Vec（2013）	无关	CBOW/Skip-Gram	相邻词	需要编码	词级负采样	词级
GloVe（2014）	无关	全局词频统计&共现概率矩阵	相邻词	需要编码	词级负采样	词级
ELMo（2018）	相关	BLSTM	相邻词	需要每层设置参数	无负采样	词级
BERT（2018）	相关	Transformer	遮	多层感知器/	短句子级负	句

XLNet（2019）	相关	Transformer-XL&自回归模型	蔽词遮蔽词	线性分类器 多层感知器/线性分类器	采样 长句子级负采样	子级句子级

通过对比分析发现，从 Word2Vec、GloVe 到 ELMo 再到 BERT、XLNet，实质是一个预训练编码由上下文无关到的静态词级向量到上下文相关的动态句子级向量的演变过程，并且将下游任务的操作逐渐转移和封装到预训练词嵌入模型的上游任务过程中去，使得下游任务的调用和实现更简单，不再需要特别复杂的编码，只需要精调参数并在预训练模型上增加一个简单的多层感知器或线性分类器即可实现多种分类预测任务。前期调研还发现，目前应用比较广泛的词嵌入模型为 Word2Vec 和 BERT，但是适用场景不同。其中 Word2Vec 比较适合于词级任务，且对词在文本中的顺序要求不高甚至是可以忽略的任务场合，如主题相关度计算、语义关系编码、文本特征聚类等。而 BERT 比较适合于句子级任务且对词的句序、句法和语义要求比较高的任务场合，如情感计算、自动问答、机器翻译等。此外，有许多研究应用已证明了 Word2Vec 的良好性能，如 Kim[24]利用 Word2Vec 预训练新闻词嵌入作为卷积神经网络的输入用于句子分类任务，在多个基准集上取得了优异结果；如 Yu 等[25]利用 Word2Vec 预训练术语嵌入来编码上下位关系属性，学习到的术语嵌入不依赖于领域且比其他嵌入具有更高的准确性。鉴于此，本文将以 Word2Vec 相关的算法模型为基础进行训练与精调以实现领域文本特征词嵌入模型的自动训练生成。

3.2 基于 Word2Vec 的词嵌入模型训练与优化方法

3.2.1 预训练方法：Skip-Gram 模型 为了使学习到的词嵌入能够有效地捕捉到低运行时复杂度下单词之间的语义关系，Word2Vec 的首创者 Mikolov 等主要设计提出了两种预训练模型：CBOW 和 Skip-Gram。CBOW 模型训练时的初始化输入是目标特征词对应的上下文相关词的词向量，输出是目标特征词的词向量。Skip-Gram 模型与 CBOW 模型的预训练测策略正好相反，初始化输入是目标特征词的词向量，输出是目标特征词对应的上下文相关词的词向量。如下面一段英文文本，若将上下文预测窗口大小设置为 4，并假设当前训练输入的目标特征词为W_t，则对应的上下文词应是该词前后的各 4 个词，见图 2。若是中文文本，则需要先分词，并去除停用词、一些自定义词及特殊符号后才能作为预测输入，见图 3。

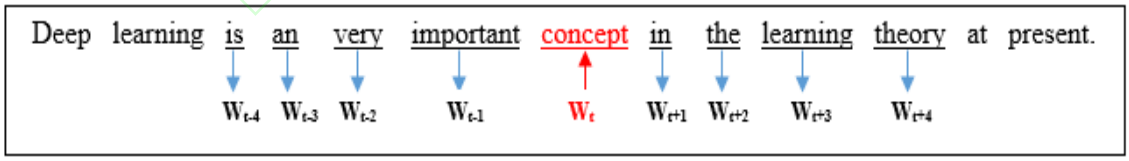

图 2Word2Vec 预训练的目标词与上下文词示例-英文文本

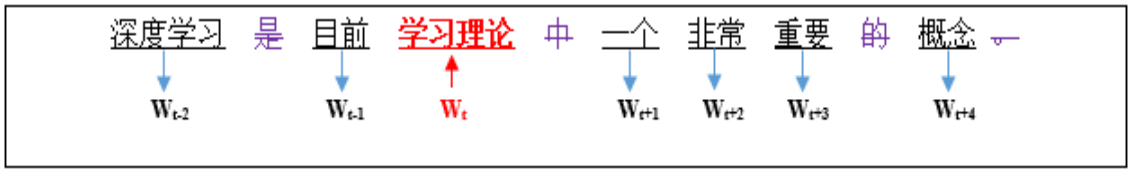

图 2 Word2Vec 预训练的目标词与上下文词示例-中文文本

前期研究发现，Skip-Gram 模型在训练时会将文本中的每个词轮流作为中心词（目标词），依次对其上下文语境范围（窗口大小）中的词进行采样，与 CBOW

相比，预测执行的次数更多，对重要的低频词的识别效果更好，因而本文选择Skip-Gram 模型作为 Word2Vec 的预训练模型。假设要预训练文本中所包含的词汇序列为$\{W_1, W_2, \cdots, W_i, \cdots, W_L\}$，$L$ 表示文本序列的长度，则 Skip-Gram 模型的预测目标是尽可能地最大化概率P_w：

$$P_w = P(W_{i-\text{window}}|W_i) \cdots P(W_{i-2}|W_i) * P(W_{i-1}|W_i) * P(W_{i+1}|W_i) * P(W_{i+2}|W_i) \cdot\cdot \cdot P(W_{i+\text{window}}|W_i)$$

其中，i应满足$t - \text{window} \leq i \leq t + \text{window} \leq L, i \neq t$，window 即为代表上下文窗口大小的常数。只要在 window 范围内的词，即便不相邻，Skip-Gram 模型也会两两组对计算其概率，因而一般情况下，window 越大，意味着模型要预测的词对就越多，预测结果相应也会越准确，但训练的时间和计算复杂度也会成倍增加。Skip-Gram 模型计算概率的核心数学公式为：

$$\mathrm{P}(W_i|W_t) = \frac{e^{U_i \cdot V_t}}{\sum_{j=1}^{V} e^{U_j \cdot V_t}}$$

其中，V_t代表W_t的输入向量，也是嵌入层矩阵中的列向量；U_j代表W_t的输出向量，也是归一化层矩阵中的行向量；V是词典的大小。从公式可以看出，Skip-gram 模型是一个中心对称模型，如果W_i在以W_t为中心词的上下文窗口大小内，反之则亦然。因而，Skip-gram 模型的预测实质是计算待预测词汇的输入向量与目标词汇的输出向量之间的余弦距离，并进行归一化输出。

3.2.2 优化加速方法：负采样算法　在实际运算中，若直接对词典中的 V 个词（常在百万级别）与预训练文本中的全部词（常在数十亿级别）两两执行余弦距离计算并进行归一化输出，则运算总量会高达数千万亿，对一般研究来说显然是一件极其耗时且难以承受的任务。因此，Mikolov 又主要设计引入了两种优化加速算法对模型的核心计算过程进行改进：分层归一化算法和负采样算法。其中分层归一化算法的实质是将原来对 V 个词的复杂归一化概率转化为 $\log V$ 个词的二分类条件概率乘积，使得计算的总复杂度从 V 降低到了 $\log V$。但分层归一化算法存在词与词之间概率计算的耦合性很强，对低频词的遍历路径比较长等缺点，因而不如负采样算法应用广泛。

综上研究，本文主要采用负采样算法对 Skip-Gram 模型进行优化加速。负采样算法是在经典的噪音对比评估算法的基础上提出的一个改进和简化版本，其核心思想是将预训练文本中某些词汇序列的中心词W_t基于一定策略替换为Neg个其他不同的词（即所谓的负例词），假设为$F_n, n = 1,2, \cdots, \mathrm{Neg}$，则$F_n$与原$W_t$的上下文词序列$\mathrm{Context}(W_t)$就构成了$\mathrm{Neg}$个虚拟的负例样本，然后对正例样本和负例样本执行二元逻辑回归计算以求解模型参数。具体的负采样策略如下：假设词典大小为 V，将长度为 L 的文本序列按词典大小及词频分成 V 块，每个文本块对应词典中的一个词w，由于每个词的词频不一样，因而每个词对应的文本长度也会不一样，相应高频词的文本长度较长，低频词的文本长度较短。Word2Vec 中定义的每个词w的文本长度$\mathrm{Length}(w)$的计算公式如下：

$$\mathrm{Length}(w) = \frac{\mathrm{count}(w)^{3/4}}{\sum_{v \in V} \mathrm{count}(v)^{3/4}}$$

但同时，由于该过程并不是均分，为了确保每个词对应的文本长度都能被成功地划分入相应的文本块，在采样前，还需首先将长度为 L 的文本序列均分为 M 份，只要$M \geq V$，那么 M 份中的每一份都将必然落入某个词对应的文本块中。采样时，只需要从上述 M 份中随机选取N个节点，获取每一个节点所属的文本块所对应的词就是最终采样到的负例词。有了负例词就可以对正例样本的中心词进

行替换，按一定比例自动生成负例样本。经前期研究，M 的取值一般为 10^8。

3.2.3 模型求解方法：随机梯度下降法　基于上述负采样策略，基于 Skip-Gram 的 Word2Vec 模型的最终优化目标就转变为：估计一个值R，既能最大化正例样本的概率（使得输出正例的可能性 r 无限接近于 1）又能同时最小化负例样本的概率（使得输出负例的可能性$1-r$无限接近于 0）。常用的是数学中最大似然估计方法中的对数似然函数来表达目标函数公式：

$$R=\sum_{n=0}^{\mathrm{Neg}} r\log\left(\sigma\left(X_{W_t}Y_{F_n}\right)\right)+(1-r)\log\left(1-\sigma\left(X_{W_t}Y_{F_n}\right)\right)$$

其中，$\sigma\left(X_{W_t}Y_{F_n}\right)$和$1-\sigma\left(X_{W_t}Y_{F_n}\right)$分别代表我们对正例和负例的期望值；$X_{W_t}$和$Y_{F_n}$分别表示正例样本和负例样本的模型参数，这两个参数就是我们最终的训练求解目标。首先构造梯度（参数的偏置向量）表达式以确定损失函数，然后采用随机梯度下降法（Stochastic Gradient Descent，SGD）[25]，每次随机选取一组样本进行学习并基于反向传播算法来更新参数权重和梯度，重复该过程，通过对全部样本进行不断迭代更新从而逐步实现对参数的求解。具体求解流程如下：

1) 算法输入：经过分词与规范化处理的预训练领域语料文本；Skip-Gram 模型对应的方法 skipGram，词向量维度大小为 layerSize，上下文窗口大小为 window，随机梯度下降迭代的初始步长（学习速率）为 alpha，负采样的样本个数为 negative 等，见表 2。这些输入一般又被称为 Word2Vec 的超参数，可根据具体需求不断进行实验调整以提高输出的词嵌入模型的性能。

2) 算法输出：全部词的词向量X_{W_t}，每个词对应的模型参数Y_{F_n}。

3) 运算过程：①初始化定义各个参数及X_{W_t}和Y_{F_n}。②读取并对预训练语料中的每组训练样本（正例）$(\mathrm{Context}(W_t),W_t)$，基于负采样算法策略采样出 negative 个负例中心词F_n，生成 negative 个负例样本$(\mathrm{Context}(W_t),F_n)$。③调用 skipGram 方法，采用随机梯度下降法，每次随机选取一组样本进行学习，然后不断迭代计算求解更新X_{W_t}和Y_{F_n}，直到梯度收敛。

表 2 Word2Vec 模型训练的主要输入参数及常规取值

参数名称	含义描述	常规取值
layerSize	特征词向量的维度（特征个数）	[100,500]间的整数
window	上下文窗口大小	[3,10]间的整数
alpha	迭代的初始步长（学习速率）	0.025
negative	负采样的样本个数	[5,20]间的整数
sample	高频词亚采样的阈值	[1e-3,1e-5]间的指数
isCbow	是否使用 CBOW 模型	false，表示使用 *Skip-Gram*
Exp_Size	逻辑函数运算的指数值	1000
Max_Exp	最大允许的指数范围	6，表示[1e-6,1e6]间的指数
hs	是否使用分层归一化算法	false，表示使用负采样算法

关于表 2 参数的一些详细说明：

对于 layerSize，一般来说，在一定范围内，词向量维度设置的越高，表明要训练出的特征就越多，词嵌入模型的精准度就会越好，但同时训练和计算的时空复杂度也会相应增加。但若是无限制地增加 layerSize，精准度并不会无限制的增加反而会增加训练运算负荷。结合前人研究，建议取值为 200 或 300。对于 window，Skip-Gram 模型的建议取值为 10，CBOW 为 5。对于 negative，适当地增加负样

本的数量，会避免模型对正例样本的过度拟合，从而提高模型的精准度，但目前并没有较佳建议值，一般取值为 5～20 之间的整数。对于 sample，当训练样本集比较大的时候，采用高频词亚采样能够提高训练速度和确保模型精度，建议取值为 1e-3 到 1e-5 间的指数，常用的是 1e-3。对于 Exp_Size 和 Max_Exp，由于 Word2Vec 模型在训练与计算时使用了很多逻辑回归运算公式（指数函数），因此，需要预先给定指数值以限定模型的运算空间和提升训练的速度。这两个值越大，词嵌入模型的精度会越高，但内存占用也会很多，过度增加会引起内存溢出，因而 Exp_Size 建议取值为 1000，Max_Exp 为 6，表示最大允许指数范围为[1e-6,1e6]。

4 领域本体概念术语的自动抽取模型构建研究

4.1 基于 IOB 标签格式的术语标注

基于前期研究，本文将概念术语的抽取问题转换为在预定义的一组文本序列上的联合分类任务。实现这一任务的起始仍然需要一批训练样本集，主要是对分词后的领域文本中的术语词汇进行分类序列标注，用于明确定义不同类型术语的边界，以处理领域概念术语中词汇的长度不一、中英文字符混杂、中英文字符和数字混杂等问题。

目前序列标注中常见的标签格式有 IO、IOE、IOB、IOBS、IOBES 等，不同的标签格式标记的侧重点和粒度因需而异有所不同，但在整个标签体系中，具有大致相同的含义：B 表示术语的首字词（Begin）；I 表示术语的中间字词（Inside）；E 表示术语的最后一个字符（End）；O 表示其他非术语词汇（Outside）；S 表示由单字词构成的术语（Single）。其中最常用的为 IOB 和 IOBES，IOB 通常适合一般粒度的标注任务，处理过程相对简单；IOBES 适合更为细粒度的标注任务，处理过程比较复杂。如大多命名实体识别模型[26-28]主要用 IOB，常见分词和词性标注工具如 NLPIR、HanLP、Jieba 等主要用 IOBES 进行语料标注与训练。

鉴于以上研究，本文选择相对简单也适合 NER 任务的 IOB 格式进行领域术语标注，并根据实际调研分析，把中文文本中常见的领域概念术语形式划分为以下几类进行分类标注，见表 3。经过分类标注后的文本，将作为 BLSTM-CRF 模型的输入，同时文本中的每个术语词汇将根据上一章节基于 Word2Vec 生成的领域文本特征词嵌入模型的计算后映射到相应维度的实值向量 D。然后，通过将术语词汇的自身向量与前后向量的上下文窗口连接，形成对该术语的最终测量。

表 3 基于 IOB 标签格式的中文概念术语分类标注

分类描述	IOB 标记	示例
纯中文	B-zh, I-zh	温跃层
纯英文	B-en, I-en	DNA
中英文混合	B-ze, I-ze	ABS 规范
数字及中文混合	B-dz, I-dz	3 维空盒子
数字及英文混合	B-de,I-de	CO2
数字及中英文混合	B-dze, I-dze	DZS2 孔
其他非概念术语词	O	一些泛指词、停用词等

4.2 基于自注意力机制的 BLSTM-CRF 模型构建

传统的深度学习算法（LSTM），主要是在深度神经网络中新增了一个包含了三个门（输入/遗忘/输出）的记忆模块，用以捕获事件中的长距离依赖关系。但 LSTM 通常只学习了文本序列的前序信息（上文信息），并没有将后序信息（下文信息）也纳入学习体系，学习到的仍是不完整的文本语义特征。BLSTM 又称

双向 LSTM，是相关学者以 LSTM 为基础进行的改进和演化，实质是两个 LSTM 的上下叠加，具体是指在同一时刻网络即计算文本的前序特征，又计算后序特征，能够组合生成更为完整的文本语义表示。已有研究表明，BLSTM 是一种更加高效的特征学习方式，能同时建模长文本序列的上下文特征，提高模型的精度。如 Chalapathy 等[28]利用 BLSTM 对临床医学概念进行提取和分类，取得了比 HMM、CRF 等 CE 算法更优的结果；如马建红等[29]利用 BLSTM 实现了新能源汽车领域术语抽取，试验结果表明模型精度比 LSTM、RNN、CRF 等更高。因而，基于以上研究，为了更有效地捕获和利用长文本序列中的上下文信息，以学习和挖掘到更完整的隐含语义特征，本文采用 BLSTM 算法进行特征学习。

4.2.1 模型的输入层和 Dropout 优化策略　BLSTM 的输入层实质是输入文本的向量表示层。该层主要是加载预训练的领域文本特征词嵌入模型并将其初始化为词嵌入矩阵，通过查找法将输入文本序列中的每个词W_t映射表示为向量X_{W_t}，向量的大小即为词嵌入模型的维度。此外，在进入下一层之前，通常还需要预先设置 dropout 比率以避免训练出现过拟合现象。过拟合现象是指算法模型的复杂度过高，过于依赖训练样本的某些局部特征，对训练样本具有非常好的预测性能，但对实际测试样本的预测性能却较差的情况。Dropout[30]是指在每个训练批次中，随机让指定比例的特征检测器（神经网络中的隐藏层节点）不工作，处于休眠状态，下一个批次会唤醒，但相应还会有其他节点再次休眠。该方法已被前人证明，可以有效地减少特征检测器之间的相互作用和避免过拟合问题，尤其在小批量训练样本中效果显著，已成为当前训练深度神经网络模型的一种必要优化策略。Dropout 比率一般为[0,1]之间的小数，当取值为 0.5 时，随机生成的网络结构最多，实际取值还可以根据训练样本的情况进行提高或降低。

4.2.2 模型的隐藏层和自注意力机制实现　BLSTM 的隐藏层实质是上下文向量拼接层。该层中，对于任何给定的文本序列中 t 时刻目标词汇W_t的向量表示X_{W_t}，网络既考虑前序信息$\overleftarrow{h_{W_t}}$，又考虑后序信息$\overrightarrow{h_{W_t}}$，通过将前后序信息对应的向量拼接为$h_{W_t} = [\overleftarrow{h_{W_t}}, \overrightarrow{h_{W_t}}]$来进行隐含特征计算。整个 BLSTM 深度神经的所有隐藏层都将基于h_{W_t}进行隐含特征计算。最终对 t 时刻目标单词W_t的预测是将其自身向量X_{W_t}和前后序信息（上下文窗口）的向量h_{W_t}全部连接在一起进行的。此外，通常考虑的上下文信息都是位于目标单词的上下文窗口范围内，窗口长度有限，受运算复杂度及计算能力限制等，并不能无限制增大。而对于领域文本来说，其他非目标词上下文窗口内的重要关键词也很可能对目标词的预测计算存在重大启示作用。因此，本文尝试在 BLSTM 的隐藏层引入自注意力模型（Self-Attentation Model）[31]来实现特征扩展。自注意力模型是近年来深度学习界提出的用于处理长序列数据的重要概念机制，已被 Google 应用在其发布的 BERT 模型的 transformer 编码器中，并显著提高了模型精度。该模型的实质是模拟人脑在特定时刻会将注意力集中在特定关键事物而常忽略其他非关键事物的专注特性，可以对关键事物特征分配较多注意力，其他则较少或不分配。基于前期调研发现，领域本体概念术语一般为（复合）名词短语、动名词短语（动宾关系）、名动词短语（主谓关系）、形容词+名词短语（定中关系）等，因而本文实现的自注意力机制是在预处理过程中对领域文本分词的同时进行词性标注，对具有上述词性的关键词进行重点关注，在特征计算时为其分配更高的权重系数，以凸显其重要作用。自注意力机制的输入为 BLSTM 隐藏层的状态向量，输出为经过自注意力机制打分后的向量。假设 t 时刻的输入为h_t，查询目标为 q，打分函数为$Score_t$，L 为原始输入文本序列中要计算的词汇的个数，则输出为：

$$\text{Attention}(q)_t = \text{Score}_t h_t = \text{Softmax}\left(\frac{\text{Sim}_t}{\sqrt{d_k}}\right) h_t = \frac{e^{\text{Sim}_t}}{(\sum_{s=1}^{L} e^{\text{Sim}_s}) \cdot \sqrt{d_k}} h_t$$

其中，Softmax是打分函数的数学公式，实质是一个计算相似度并求指数运算的过程，$\sqrt{d_k}$是自注意力机制的原作者 Vaswani 等设计提出的一个优化参数，又称多头注意，是指将原向量模型的维度d_v再线性映射到不同空间 h 次，然后再进行注意力计算，以避免在d_v维度下Sim_t函数内积过大，造成Softmax指数运算结果非 1 即 0 的问题。一般，$d_v = 512, h = 8, d_k = d_v/h = 64$。

4.2.3 模型的输出层和 CRF-Viterbi 解码　BLSTM 默认的输出层是一个被命名为 tanh 的激活函数，尽管该函数已在许多情况下被证明是有效的，但它无法使用机器学习中最常用的动态规划算法——维特比算法（Viterbi Algorithm）[32]进行联合解码，不能很好地识别和处理不同分类标签之间的复合结构、嵌套结构等复杂依赖关系。如“3 维空盒子”之类的概念术语，既属于【数字及中文混合类】，又可能属于【数字及中文混合类】和【纯中文类】的复合体，在原始 BLSTM 的输出模式下，会将其拆分识别为“3 维”和“空盒子”两个短术语，而不是作为一个完整的长术语来考虑。或者会错误识别，导致一个术语的不同位置字符被拆分出现在不同的分类预测标记中，如$[\text{B}-\text{zh}, \text{I}-\text{en}, \text{O}, \text{O}, \cdots]$这种明显不合法的预测结果。因而，目前得到较多认可的对 BLSTM 的另一个改进是在 BLSTM 的原输出层后再添加一个条件随机场层，以实现最佳顺序的解码，这也是 BLSTM-CRF 模型的由来。CRF 作为经典的自然语言处理模型，已被成功应用于传统的命名实体识别、序列标注、中文分词等任务且效果良好。BLSTM-CRF 模型中 CRF 层实现的核心任务是：进行线性变换，计算（优化）损失，实现维特比解码。具体流程是：首先，将 BLSTM 的原输出特征通过线性变换的方法转换为一个维度为[训练批次大小，文本序列长度，总标注类别数]的矩阵（张量）。其次，将矩阵、相应的标记序列、每一个序列的长度等作为参数输入到预定义的损失函数中进行计算，并得到一个转移矩阵。损失函数是目前用来衡量机器学习模型好坏的一个关键指标，值越大表明模型预测误差越大，因而需要根据损失值不断更新模型参数及函数，在参数空间内寻找最优解，使得损失函数的值越小越好。本章使用的 CRF 损失计算函数是一个最大似然估计函数，在主流深度学习框架中都有定义。最后，将第一步的线性变换矩阵和第二步得到的转移矩阵等作为参数输入维特比算法进行解码，获得维特比序列和维特比打分值，并计算损失值。在整个训练过程中，会通过随机梯度下降法来不断优化损失，提升模型精度。

5 领域本体概念的自动获取实验与结果分析

5.1 实验数据集的获取与预处理

基于相关工作需要，本文以资源环境学科领域为例展开试验研究。由于目前网络上并没有公开可获取到的资源环境学科领域的中文标准语料集，本文需要自行构造语料集。为了保证实验所用的领域语料既具有新颖性、前沿性又具有一定的权威性、专业性，本文同时采集了领域科技动态类语料和以期刊/会议论文为主的学术文献类语料。其中，领域科技动态类语料主要通过将中国科学院兰州文献情报中心服务网站的环境与发展动态栏目[33]及专题信息网站的资源环境科技发展态势分析平台的资源环境快报栏目[34]作为采集源，设计 WebSpider 对其内容信息进行自动监测和采集抽取。这些栏目的信息都是经过领域情报人员编译或遴选编辑的最新热点信息，能够确保领域的相关性和专业性。领域科学文献语料的获取主要通过单位已订购的 Web of Science 数据库，选择中国科学引文数据库（CSCD），以预定义的资源环境学科领域主题检索式进行高级检索，时间跨度为

2010—2019，基于数据库提供的 Web Services API 编写数据调用程序和 XML 解析程序对期刊/会议论文相关的元数据进行自动采集和解析抽取。最终共获取到有效的领域科技动态类语料 4371 篇，中文学术文献类语料 398975 篇，共 403346 条数据记录。接下来，首先，将每篇科技资讯的标题和正文内容（去除全部 HTML 标签），学术文献的标题和摘要依次以换行符（“\r\n”）为分隔输出和整合在同一个 txt 文档中，即为初始语料集 Resource_Corpus。其次，将 398975 篇中文学术文献中的中英文作者关键词提取出来，经过一系列去重、计算词频、删除通用词等操作，生成领域基础知识词典 Resource_Dic，共得到有效词组 501755 个，见表 4。然后，基于 Resource_Dic 和 HanLP 提供的 SP 基础分词模型和标注框架，增量训练出经过领域泛化的 SP 词法分析模型，并对 Resource_Corpus 进行分词、词性标注和 IOB 标注，见表 5。此外，我们将标注过的数据按总条数 8:1:1 的比例切分开，其中 80%作为训练集，剩下各 10%分别作为测试集和验证集。

表 4 资源环境领域基础知识词典示例

中文关键词	英文关键词	词频	词性
数值模拟	Numerical simulation	4439	名动词/nv（主谓关系）
气候变化	Climate change	3935	名动词/nv（主谓关系）
产量	Yield	2865	名词/n
水资源	Water resource	2827	名词/n
重金属	Heavymetals	2614	名词/n
地球化学	Geochemistry	2571	名词/n（并列关系）
降水	Precipitation	2094	动词/v
影响因素	Influencing factors	2081	名词/v
土壤	Soil	1997	名词/n
GIS	GIS	1598	外文名词/nx

表 5 资源环境领域文本分词及标注示例

分词	厄	尔	尼	诺		期	间		北	美		生	物	圈		二	氧	化	碳
词性	专有名词（nz）					方位词（f）			地名（ns）			名词（n）				名词(n)			
标注	B-zh	I-zh	I-zh	I-zh		0	0		0	0		B-zh	I-zh	I-zh		B-zh	I-zh	I-zh	I-zh
分词	（		C	0	2		）		吸	收	量		显	著		增	加		。
词性			英文专有名词（nx）						动名词（vn）				形容词（a）			动词（v）			
标注	0		B-de	I-de	I-de		0		B-zh	I-zh	I-zh		0	0		0	0		0

5.2 实验方法、工具与关键过程

本节的实验环境为 Linux 操作系统，版本为 CentOS 7.5，64 位，32GB 内存。在领域文本特征词嵌入模型的自动生成实验环节，首先在 GitHub 上下载 Word2Vec 的 Python 源码进行安装部署；其次对 Word2Vec 进行调参，如表 2 及参数说明所述，主要设置特征词向量维度为 300，上下文窗口大小为 10，负采样的样本个数为20，高频词亚采样的阈值为le-3，其他保持默认；然后调用 Word2Vec 的 Skip-Gram 模型，将领域基础知识词典中的词生成的文本作为模型输入，使用负采样算法和随机梯度下降法进行训练，训练结果即为资源环境领域文本特征词嵌入模型 Resource_Dic.Model。在领域本体概念自动抽取模型的构建实验环节，主要使用 Theano[35]深度学习工具包，首先下载和安装 Python 的集成开发环境 Anaconda3，并在其中安装和搭建 Theano 运行环境；其次调用 Theano 相关组件和函数，构建和开发基于自注意力机制的 BLSTM-CRF 深度学习算法网络程序；接着对算法网络进行调参，包括设置 BLSTM 隐藏层节点数量为 100，上下文窗口大小为 10，词嵌入维度为 300，标注模式为 IOB，学习方法为 SGD，解码方法为 CRF，Dropout 比率为 0.5，训练停止设置为 100 个 epochs 以避免过拟合问题。然后加载 Resource_Dic.Model 作为词向量查找表，预先准备的训练集、测试集和验证集作为输入，进行模型训练和优化，最终保存在验证集上具有最佳性能的参数及模型。

5.3 实验模型的评估与结果分析

本文对实验模型的评估采用著名的计算自然语言处理会议 CoNLL 提供的 conlleval 评分脚本，主要评估标准为当前应用比较广泛的准确率、召回率和 F1 值。由于训练比较耗时，主要进行了三组模型的实验，见表 6。

表 6 资源环境领域概念自动抽取模型的实验结果

实验模型	准确率	召回率	*F*1
自注意力机制+BLSTM-CRF	0.896	0.835	0.864
BLSTM-CRF	0.834	0.798	0.816
LSTM-CRF	0.782	0.751	0.767

从实验结果可以看出，基于自注意力机制的 BLSTM-CRF 模型相对于其他常用模型来说，概念术语抽取效果明显有了较大的提升。原因主要在于，除了借鉴已有 BLSTM 模型的优点外，如使用 CRF-Viterbi 解码能够处理原来 BLSTM 难以解决的不用分类标签之间的复杂依赖关系外，本文设计的自注意力机制还可以在隐藏层进行特征捕获计算时为具有重要词性的概念术语赋予较高的权重系数，从而使模型的注意力集中在更为关键的特征上，降低了无关特征的干扰作用。此外，与以往的需要构造复杂特征模板的 CRF 模型相比，BLSTM-CRF 模型并没有使用太多手动设计的抽取特征，预训练的领域文本特征词嵌入模型是完全利用 Word2Vec 从领域纯文本数据中无监督学习，自动抽取模型是从基于 IOB 格式的标准文本中以端到端的识别模式进行训练学习，因而对于领域新概念或新术语词汇的识别也具有一定的健壮性和可扩展性。

6 结论与未来工作

研究表明，本文以无监督的学习方法和端到端的识别模式为理论技术基础，构建的基于自注意力机制的 BLSTM-CRF 模型能够不依赖于人工设计抽取特征

而实现对领域文本复杂隐含特征的自动挖掘，方法流程具有较好的领域通用性和可移植性，一定程度上解决了领域本体概念术语获取的共性问题。在实践应用方面，本模型已有效实现资源环境领域本体概念术语的自动识别抽取，并支持可快速移植于其他领域的概念获取，将为不同领域本体自动化构建中领域概念的生成提供可用方法工具。存在的局限是，虽然模型精度有了较大提升，但尚未达到0.9 以上的理想峰值。未来为了进一步提高模型的精度，值得考虑改进的地方还有很多。如在领域文本特征词嵌入模型的预训练方面，可考虑用 Glove、Bert 等模型进行训练和精调测试，进一步选择更优秀的预训练模型；虽然已经有了 50 多万的术语作为领域词典，但对于一个大领域来说可能还不够，仍可考虑扩展预训练的词汇样本量；调整特征词嵌入的维度等。在 BLSTM-CRF 自动抽取模型的构建方面，如用于捕获特征的隐藏层节点的数量，词嵌入的维度，输入文本的标注分类及标注模式等都可以通过不断测试而调优。□

参考文献

[1]张晓林. 颠覆性变革与后图书馆时代——推动知识服务的供给侧结构性改革[J]. 中国图书馆学报，2018，44 (1)：4-16.

[2]刘柏嵩. 中文领域本体自动构建理论与应用研究[M]. 杭州：浙江大学出版社，2014.

[3]KANZAKI K, BOND F, TOMURO N, et al. Extraction of attribute concepts from Japanese adjectives [J]. Language Resources and Evaluation, 2008.

[4]ABACHA A B, ZWEIGENBAUM P. Automatic extraction of semantic relations between medical entities: a rule based approach[J]. Journal of Biomedical Semantics, 2011, 5(2): 1-11.

[5]GUPTA S, MANNING C. SPIED: Stanford pattern based information extraction and diagnostics[C]. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, USA, 2014: 38-44.

[6]王昊, 苏新宁. 基于模式匹配的中文通用本体概念抽取模型[J]. 情报理论与实践, 2008(2):292-297.

[7]BUN K K, ISHIZUKA M. Topic extraction from news archive using TF*PDF algorithm[C].Proceedings of the Third International Conference on Web Information Systems Engineering, 2002: 73-82.

[8]WEI X L, SUN Y, ZHANG S K, et al. Ontological concept extraction method based on maximum entropy model[J]. Computer Engineering, 2009, 35(24): 114-116.

[9]唐晓波,胡华.中文 UGC 信息源的本体概念抽取研究[J]. 现代图书情报技术, 2014, 30(5): 41-49.

[10]颜端武, 李兰彬, 曲美娟. 基于 N-gram 复合分词的领域概念自动获取方法研究[J]. 情报理论与实践, 2014, 37(2):122-126.

[11]樊梦佳, 段东圣, 杜翠兰, 等. 统计与规则相融合的领域术语抽取算法[J]. 计算机应用研究, 2016, 33 (8): 2282-2285，2306.

[12]余凡, 楼雯. 领域概念的三层递进筛选方法研究[J]. 现代图书情报技术, 2015, 31(4): 26-33.

[13]WU B, OUYANG L B. A method of domain compound concept extraction based on multilevel filter[J]. Computer and Information Technology, 2014:2292-2296.

[14]HABIB B M, KEULEN M V, ZHU Z. Concept extraction challenge: university of Twente at MSM2013[C]. Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts', Germany, 2013: 17-20.

[15]BRUIJN B D, CHERRY C, KIRITCHENKO S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 557-562.

[16]CHI C Y, ZHANG Y. Information extraction from Chinese papers based on hidden Markov model[J].Advanced Materials Research,2014, 846-847: 1291-1294.

[17]ZHANG W, YOSHIDA T, TANG X J. Using ontology to improve precision of terminology extraction from documents[J].Expert Systems with Applications, 2009, 36(5): 9333-9339.

[18]NASTASE V, STRUBE M. Transforming Wikipedia into a large scale multilingual concept network[J]. Artificial Intelligence,2013,194: 62-85.

[19]MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL].[2013-09-07].https://arxiv.org/abs/1301.3781.
[20]PENNINGTON J, SOCHER R, MANNING C D. GloVe: global vectors for word representation[DB/OL]. [2018-12-29]. https://nlp.stanford.edu/projects/glove/.
[21]PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C].Proceedings of NAACL-HLT 2018,New Orleans, Louisiana, 2018: 2227-2237.
[22]DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL].[2019-05-25].https://arxiv.org/abs/1810.04805.
[23]YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding[EB/OL].[2019-06-19].https://arxiv.org/abs/1906.08237.
[24]KIM Y. Convolutional neural networks for sentence classification [EB/OL]. [2014-09-03]. https://arxiv.org/abs/1408.5882.
[25]BOTTOU L.Stochastic gradient descent tricks[C].Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg,2012: 421-436.
[26]YU Z, WANG H X, LI X M, et al. Learning term embeddings for hypernymy identification[C]. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, 2015: 1390-1397.
[27]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(1):116-122.
[28]CHALAPATHY R, BORZESHI E Z, PICCARDI M. Bidirectional LSTM-CRF for clinical concept extraction[EB/OL]. [2016-10-19]. https://arxiv.org/abs/1610.05858.
[29]马建红, 张亚梅, 姚爽, 等. 基于 BLSTM_ attention_CRF 模型的新能源汽车领域术语抽取[J]. 计算机应用研究, 2019(5):1385-1389.
[30]SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
[31]VASWANI A, SHAZEER N, PARMAR N. Attention is all you need[EB/OL].[2017-12-06].https://arxiv.org/abs/1706.03762.
[32]李航. 统计学习方法[M].北京：清华大学出版社, 2012.
[33]环境与发展动态[DB/OL].[2019-07-22].http://www.llas.cas.cn/xwzx/kxxw/.
[34]资源环境科技发展态势分析平台[DB/OL].[2019-07-22].http://resp.llas.ac.cn/.
[35]Theano[DB/OL].[2019-06-28].http://deeplearning.net/software/theano/.

作者简介：王思丽（ORCID: 0000-0002-2126-3462），女，1985 年生，博士生，馆员。研究方向：知识发现与知识组织，知识计算与知识挖掘。**祝忠明**（ORCID:0000-0002-2365-3050），男，1968 年生，研究馆员，博士生导师。研究方向：知识发现与知识组织，知识管理系统建设。**刘巍**（ORCID:0000-0001-6387-1709），男，1980 年生，副研究馆员，硕士生导师。研究方向：知识计算与知识挖掘。**杨恒**，男，1992 年生，硕士，助理馆员。研究方向：分布式大数据系统建设。
作者贡献声明：王思丽，设计研究方案，构建抽取模型，论文起草、撰写与修订。**祝忠明**，提出整体研究思路，指导整体研究过程。**刘巍**，参与抽取模型构建，指导关键技术实现。**杨恒**，参与词嵌入模型训练，概念抽取实验。
收稿日期：2019-09-09