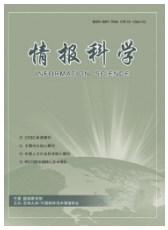

情报科学
Information Science
ISSN 1007-7634,CN 22-1264/G2

《情报科学》网络首发论文

题目：基于 BERT 的领域本体分类关系自动识别研究
作者：王思丽，杨恒，祝忠明，刘巍
收稿日期：2020-04-17
网络首发日期：2020-10-23
引用格式：王思丽，杨恒，祝忠明，刘巍．基于 BERT 的领域本体分类关系自动识别研究[J/OL]．情报科学.
https://kns.cnki.net/kcms/detail/22.1264.G2.20201022.1502.008.html

网络首发时间：2020-10-23 11:32:53
网络首发地址：https://kns.cnki.net/kcms/detail/22.1264.G2.20201022.1502.008.html

基于BERT的领域本体分类关系自动识别研究

王思丽[1,2]，杨　恒[1]，祝忠明[1]，刘　巍[1]

（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.中国科学院大学，北京 100049）

摘　要：【目的/意义】实现对领域本体分类关系的自动学习识别，解决领域本体知识框架结构体系的自动化构建问题。【方法/过程】通过对领域本体分类关系自动识别的国内外研究现状及存在问题进行分析总结，以当前开源的先进的深度学习文本预训练模型BERT为基础，研究构建了基于BERT的领域本体分类关系自动识别模型，并以资源环境学科领域为例进行了实验研究和评估分析。【结果/结论】模型能够实现对领域本体分类关系的自动识别，识别方法和流程具有极大地通用性和可移植性，识别精度比传统方法有了较大提升。但由于受分类标注语料的质量限制，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；分类关系识别；分类标注；BERT

Research on Automatic Identification of Domain Ontology Classification Relations Based on BERT

WANG Si-li[1,2]，*YANG Heng*[1]，*ZHU Zhong-ming*[1]，*LIU Wei*[1]，

(1.Literature and Information Center of Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000*, China;* 2.*University of Chinese Academy of Sciences, Beijing* 100049*, China)*

Abstract: [Purpose/significance] Realize the automatic learning and identification of domain ontology classification relations, and solve the problem of the automatic construction of domain ontology knowledge framework structure. [Method/process] Through the analysis and summary of the research status and existing problems of automatic recognition of domain ontology classification relationship at home and abroad, the paper studies and builds the domain ontology classification relationship automatic recognition model based on the open source and advanced deep learning pre-trained models BERT. Then, the paper takes the field of resources and environment as an example to conduct experimental research and evaluation analysis. [Result/conclusion]The model can realize the automatic identification of domain ontology classification relationships. The identification methods and processes are extremely versatile and portable, and the identification accuracy is greatly improved compared to traditional methods. [Limitations]Due to the quality constraints of the classified annotation corpus, the model accuracy has not yet reached its peak and needs to be further optimized and improved.

Keywords: Deep learning; Domain ontology; Classification relation identification; Classification annotation; BERT

1 引　言

领域本体作为一种基于概念和概念间语义关系来表示和描述特定领域知识结构体系的重要工具模型，已成为领域知识挖掘和语义分析计算不可或缺的基础要素，被认为是大数据环境下解决“信息和知识孤岛”问题的最佳实践之一[1]。笔者所在团队前期已对领域本体自动化构建的前序

收稿日期：2020-04-17

基金项目：国家科技部重点研发计划课题“应对气候变化科学数据与知识集成共享平台建设”（2018YFC1509007）、中国科学院西部之光项目“开放学术资源的情景化组织与服务研究”（Y9AX011001）、中国科学院文献情报创新能力建设项目“基于深度学习的领域本体自动构建方法研究”（Y8AJ012005）

作者简介：王思丽（1985-），女，河南南阳人，博士研究生，馆员，主要从事知识发现与知识组织研究.

任务——领域本体概念的自动获取方法[2]进行了深入研究，本文属于其后序任务。领域本体分类关系是指领域本体中的每一个概念术语都必将隶属于领域本体分类结构中的某一个或多个实体类或其子类，由此构成的知识层次关系即为分类关系。领域本体分类关系的自动识别是指基于一定的策略/模型自动分析、理解领域概念术语的内涵和外延并将其自动判别、划分、映射到预定义的领域本体主题分类结构或知识框架中去的过程。本文通过对领域本体分类关系识别的国内外研究现状及存在问题进行分析总结，以当前开源的先进的BERT模型为基础，对其核心技术原理进行剖析与改进应用，研究构建了基于BERT的领域本体分类关系自动识别模型，并进行了实验验证与评估分析，旨在有效解决现有领域本体分类关系识别过程中对人工提取特征的过度依赖、难以充分挖掘识别复杂隐含关系、模型可移植性差和自动化程度低、分类识别精度欠佳等问题。

2 国内外研究现状述评

目前国内外常见的领域本体分类关系识别方法主要有基于人工的方法、基于模式规则的方法、基于机器学习/深度学习的方法等。

基于人工的方法主要是指借助于领域专家的专业背景知识和专家个人对概念术语的主观理解分析，手动对所有概念术语关系由高到低、由根到叶依次进行判别、分类的过程。如国外大型通用领域本体WordNet、OpenGalen[3]等都曾主要依靠人工，耗费大量人力、物力成本建成。如2012年，由北京理工大学王庆林等主持构建的气候变化领域本体[4]，建设初期也曾主要依赖领域专家进行领域术语的分类挂接。如2013年，由中国科学院文献情报中心承担构建的大型外文超级科技词表STKOS本体[5]，是将本体词表按学科领域划分为多个子本体任务，分配至各领域相关的专家团队由多人手动协同加工完成。人工的分类判别方法在通用领域或大学科领域概念分类体系下，具有一定的准确性，但面对比较小而专的学科主题领域或比较深奥/生僻/相似度较高的专业术语或概念，则容易出现主观性、倾向性错误，且实施起来需投入的生产、维护成本过高。

基于模式规则的方法主要是指以人工进行浅层的句法/词汇语义分析或已经过专家编译构建的领域词表（如WordNet、HowNet等）、分类体系（学科/行业的分类法等）、在线知识库（Freebase、Wikipedia、百度百科等）的分类结构/模式为基础，通过人工设计规则和机器编程形成一套规则化、模式化应用程序/工具进行半自动分类的过程[6-8]。该方法实质是一个最大化枚举、归纳概括或机械转换，匹配的过程，不再完全依赖人工，实施成本低且更为便捷、高效，分类过程具有可回溯性。但该方法对语言环境规则的依赖性强，将既定义的一种语言的分类模式直接应用于另一种语言环境十分不易，因而可移植性可重用性差；且该方法应用场景的局限性很大，只能覆盖较少的一部分学科/主题领域，因为已有的可直接作为参考知识源的领域分类体系大多主要集中在通用学科领域，而隶属于科技领域的部分又主要集中在医学、地学等少数自然科学领域，对于没有可参考的或可参考知识源中不存在的分类关系的识别则无能为力。

基于机器学习/深度学习的方法主要是将分类关系识别问题转化为一个二分类、多分类、多标签分类、层次多标签分类等任务模式之一进行自动化处理，其核心流程如图1所示。首先，需要根据应用场景、任务目标等确定分类任务模式；其次，由于不同分类任务模式所依赖的语料集格式并不相同，需根据已确定的分类任务模式获取和准备一个具有大数据量的、高质量的、经过规范化预处理的有监督分类标注的语料集。接着，利用有监督分类标注语料集（预）训练和生成词嵌入，即实现文本数据的数值化表示。再次，由于不同分类模型的神经网络结构及所支持的分类模式各不相同，需选择和构建合适的深度/机器学习分类算法模型，对词嵌入进行编码表示、特征提取、筛选和组合等，训练和微调出最佳分类模型；最后，利用获得的最终分类模型实现对未知文本的分类。传统机器学习分类算法模型有支持向量机、朴素贝叶斯、决策树、逻辑回归等，分类效果良好，已得到大量研究应用，一直是过去分类关系识别的主流方法。当前，随着深度学习的白热化发展，传统机器学习分类算法正逐渐被深度学习分类算法所融合和取代，几乎所有的深度学习算法模型都支持分类关系自动识别。如TextCNN[9]是在2014年由美国纽约大学的Kim提出的，被认为是最经典的深度学习文本分类模型。如FastText[10]是在2016年由美国Facebook人工智能研究院的Joulin等提出的一个有监督的快速文本分类算法工具。TextRNN[11]是在2016年由我国复旦大学的Liu等针对文本多分类任务提出的一种递归神经网络模型。此后，有许多研究者开始将文本序列中的相互依赖关系纳入CNN/RNN的特征提取和计算过程，旨在学习时能够赋予重要关注内容更高的权重以提升模型性能，如多语言分层注意力网络MHAN[12]、多标签情感分类注意力卷积神经网络AttnConvnet[13]。2018年以来，谷歌与相关研究团队合作也先后发布了与以往CNN/RNN的循环式网络架构完全不同的基于多头注意力机制和变压器编码结构的BERT[14]、XLNet[15]模型，在包括文本分类在内的多个NLP任务中都拔得头筹。2019年腾讯的Liu等还推出了一个开源的集成化分类工具NeuralClassifier[16]，集成了FastText、BERT等多种主流分类模型及预训练机制，使得用户可通过系统预设的配置文件灵活调用各种分类模型。

总的来看：

（1）基于模式规则的分类方法与基于机器学习分类方法的本质区别是，模式规则方法的依赖输入是硬核的机器指令代码，而机器学习方法的依赖输入是结构化的文本数据。

（2）传统机器学习分类方法仍主要依赖人工干预进行特征选择、提取和表示，学习输入的是人工设计好的浅层/显式数值化特征，基于特征比较实现分类识别，识别的精度取决于人选取特征的能力及当前样本的外在特征表现力，因此识

别效果难以保持稳定，模型的泛化能力较差。并且传统机器学习分类方法一般擅长研究解决的是二分类问题或多分类问题，类别间不能具有相互包含或重叠等复杂关系，即便实际应用中存在一定关系，传统机器学习分类算法在计算处理时也会经常强制忽略不计。

（3）深度学习分类方法不再需要人工进行特征提取和表示，主要依靠各种深度神经网络算法模型自动提取初级基础特征，多层学习迭代后组合为高级复杂特征，基于端到端的分类概率预测实现的分类识别，极大地解放了人力，最大程度实现了自动化，且识别的精度有了很大提升。同时，深度学习分类方法不仅能够处理二分类问题或多分类问题，还能够通过各种机制策略致力于解决处理多标签分类问题及层次多标签分类问题，更加贴合于领域实际应用场景。

（4）虽然深度学习分类算法模型已表现出比其他传统分类方法更好的性能优势，但也存在一些缺点，如算法网络的构建过程比较复杂，可解释性较差，模型训练过程比较复杂，对软硬件环境的要求较高等，因而实施起来并不是十分容易。目前仍主要处于理论研究和实验阶段，尤其对一些科研实力不是很雄厚的中小型机构单位或业务领域来说，距离真正全面投入生产应用还尚需一定时日。

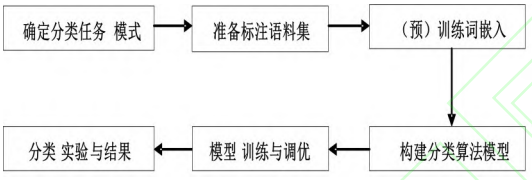

图1 机器/深度学习分类的核心流程

3 领域本体分类关系自动识别模型的构建研究

3.1 BERT的技术原理剖析

BERT是Google的Devlin等在2018年公开推出的一个基于深度双向编码器预训练的语言理解模型，已在实体识别、文本分类、自动问答等多个NLP任务实验中取得比以前方法模型更高的精度。与先前的词嵌入模型相比，BERT进一步增强了词嵌入模型的泛化能力，更加充分地描述了文本序列中字符级、单词级、句子级、甚至是句间关系的上下文特征，极大地提升了语言理解表示的能力和效率。BERT主要有三个重要的机制：

（1）变压器—注意力架构

传统的CNN/RNN在处理NLP任务时都存在相应不足，CNN的卷积操作和固定滤波器大小的限制不适合于建模长文本序列，RNN没有并行化机制，处理长文本序列很容易超出内存限制。BERT的变压器–注意力架构是一种旨在取代和改进传统CNN/RNN的全新架构，主要包含多头注意力、自注意力、位置编码等机制，可以具有更深的层数和更好的并行性。

（2）遮蔽语言模型

以往的标准条件语言模型如Word2Vec、Glove、ELMo、BLSTM等只能从左到右或从右到左进行单向或浅双向训练，因为深层双向条件训练将会导致每个单词间接地“自己看见自己”，并且模型是在多层上下文中琐碎地预测目标单词。为了获得深层双向表示模型，BERT在预训练过程中使用了一种类似“完形填空”的遮蔽语言模型，按一定比例随机遮蔽某些输入标记，然后只预测那些被遮蔽的标记，即只将被遮蔽标记对应的最终隐藏向量通过词汇表传递到输出层的归一化函数中去，而不是建模和传递整个输入。在2019年推出的最新版本中，BERT还提出了全词遮蔽技术[17]，在总遮蔽率保持不变的情况下，一次遮蔽与一个单词对应的所有标记，旨在提高预测任务的难度和增强预训练模型的泛化能力。

（3）句子级的表示模式

以往模型的编码机制学习到的大多是词级的标记特征，BERT为了支持自然语言推理、自动问答等一些需要句子间交互匹配、关系判断的更为复杂的下游任务，设计实现了句子级的表示模式。首先是句子级的负采样策略，即对任意给定的句子，原始语料序列中它的下一句即视为正例，从全部语料中随机采样一个句子即为负例，然后在句子级上进行二分类训练，即预测/判断目标句子是当前句子的下一句还是噪声。其次是句子级的嵌入表示，即词块嵌入，输入为能够在一个标记序列中明确地表示一定范围内、一个句子或句子对的连续文本，通过将标记嵌入、分段嵌入、位置嵌入拼接在一起形成，如图2所示。

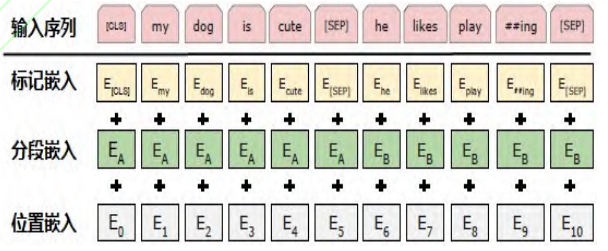

图2 BERT句子级的输入表示[14]

3.2 基于BERT的分类关系自动识别模型构建

3.2.1 BERT预训练模型的获取与准备

目前谷歌在GitHub上总共公开了9个BERT预训练模型[17]，不同预训练模型的训练策略有所不同，主要有基础版和大型号版之分；还有忽略大小写和保留大小写之分；语种方面，有英文、多语种、中文之分。其中每个模型的文件压缩包都主要包含三项内容：

①TensorFlow检查点，用于保存预训练模型与TensorFlow有关的权重参数，实际上由3个文件构成：数据、索引、元数据（TensorFlow计算图的结构）。

②vocab文件，即BERT的词汇表，包含固定数量（一般都在20000–30000左右）的单词、单字符、子字符+字符标记等，用于映射词块到单词ID。

③配置文件，用于指定或调整模型的超参数，包括drop-

out比率、隐藏层的激活函数、每个隐藏层的神经元个数、隐藏层的个数、最大允许的位置嵌入的维度、词汇表的大小等。

3.2.2 多分类标注语料的自动获取与生成

本文将领域本体分类关系自动识别问题视为一个多标签分类任务模式，分类的对象是前序任务中获取到的一堆领域本体概念，分类的类别标签可以由人工预定义或参考已有主题词表或分类体系的分类类目进行构造，最终的分类任务目标是构建和训练一个分类模型能够自动识别领域本体概念的内涵和外延，并将其准确划分到最为接近的预定义的分类类别中去。确定了分类任务模式后，即可根据该模式构造与准备所需的多标签分类标注语料。原始语料的自动获取可根据实际定义的分类类别标签分别构造相应类别的主题采集规则或检索式，利用基于Web Spider从领域专业网站自动采集或基于Web API从科学文献数据库自动收割解析内容元数据等[18]。同时，通过对通用语言理解评估GLUE基准[19]中所提供的BERT用作文本分类示例的MRPC语料集样本进行分析，确定的分类标注语料格式如下：输入类型为纯文本，一个标记序列即为文本中的一行；每行又分为两列，两列之间一般用Tab键（“\t”）隔开；第一列为类别标签（label），第二列为类别对应的主题相关的文摘内容（Text）。最终需要编写语料处理程序将原始语料每个类别按8:1:1的比例以上述格式累加输出和整合生成3个文档，分别作为训练集、验证集和测试集待用。

3.2.3 语料数据的预处理与特征表示和转换

任何模型的训练、预测都必须有一个明确的规范化的输入格式，本环节的主要功能就是加载和读取语料文件，创建数据实例，并进行标记化处理与特征转换，为BERT提供预期的输入表示。具体设计和实现方法如下：

（1）自定义数据处理模块，用于加载和读取领域分类标注语料，并生成数据实例列表，为特征提取和转换做准备。首先，需要在BERT执行分类的处理器构造类中为新增的自定义数据处理函数添加声明，包括添加声明名称和数据处理函数名称，格式为[“声明名称”: 数据处理函数名称]。其中声明名称将作为数据处理函数调用的唯一标识符，在模型执行训练或预测时，可通过在运行参数命令行中加入[--task_name=声明名称]来表示调用相应的数据处理函数。其次，对自定义的数据处理函数进行具体实现。实际上，BERT的原始模型中已经提供了用于处理序列分类数据集的数据转换器的基类。因此，实现自定义的数据处理函数需要首先继承该基类，并重载和实现基类中的一些固有方法，包括读取训练集、验证集并创建各自输入实例的集合的方法，获取数据集的所有分类标签列表的方法等。需要注意的是，不同格式语料文件的读取和处理方法不同，一旦语料文件格式发生变动，上述方法的实现应随之修改。最后，本模块返回的是一个实例列表，实例列表中的每一个实例集合都是由实例的唯一ID、第一个序列的未经标记化处理的原始文本内容（text_a）、第二个序列的未经标记化处理的原始文本内容（text_b）、类别标签（label）元素组成。

（2）自定义特征转换模块，用于读取实例列表、分类标签列表，进行标记化处理与特征转换，形成文本序列的特征表示。本模块实质就是实现如3.1小节所述的BERT句子级的嵌入表示过程，其核心是一个特征转换函数，最终将返回一个特征列表。特征列表中的每一个特征集合都主要包含四个元素：

①输入实例的ID列表：文本序列经标记化处理后每个单词在词汇表中的ID。具体生成ID的过程如下：首先，调用BERT标记器的标记化方法，对每一个句子序列实例根据词汇表和贪心原则进行基于字的分词，并按BERT规则加上[CLS]与[SEP]标记，最后得到一个经分词和标记化处理的单词列表。该过程首先会检查整个单词是否已在BERT预训练模型提供的词汇表中，对于不在词汇表中的单词，会尝试将单词首选分解为词汇表中包含的最大可能子字符，最后才分解为单个字符。其次，再次调用标记器的将标记转换为ID的方法，将得到的单词标记列表与词汇表中的索引进行匹配查找，获得每一个单词的ID。其中分类标记[CLS]和分句标记[SEP]的ID已分别固定为101和102。此外，如果实际序列实例的长度小于模型设定的最大序列长度，则缺少部分相应位置ID会被填充为0，如果大于，则超出部分会被截断。

②输入实例的遮蔽标识列表：用以标识当前标记是真实标记还是填充标记。真实标记的标识为1，填充标记的标识为0，[CLS]和[SEP]的标识也为1。

③分段ID列表：BERT支持句子对的训练，期望使用0和1来区分两个句子序列，序列A（text_a）中的每个单词分段ID全标记为0，序列B（text_b）中的每个单词分段ID全标记为1。对于文本分类任务来说，暂时不需要用到序列B，因此值都为0。

④分类标签ID列表：将分类标签列表中的类别标签名称利用枚举/遍历方法转换为索引标识，将分类标签也数值化、向量化。最常用的转换方式就是从下标0开始，第1个类别标签对应的ID为0，第2个对应为1，依次计算，第N个类别对应为N-1。

3.2.4 基于BERT的分类模型的训练与优化

经过数据处理与特征表示，可将标注语料转化为特征向量输入到BERT模型中，从而进一步实现对BERT预训练模型的微调与优化，构建基于BERT的领域本体分类关系自动识别模型。本环节总体实现流程如图3所示，具体包含以下几个重要步骤：

（1）Step1：准备模型的超参数。其中需要重点调整的参数有最大序列长度、训练批次大小、学习速率、训练周期数等。同等条件下，最大序列长度的值越大，表明模型输入计算的特征将越多，模型的精度也会得到大幅度提升，但训练所需的计算开销也会成比例增大。训练批次大小也非常重要，深度学习通常将训练集分成多个小批量样本，一次迭代一个，通过各种优化算法计算损失并更新权重。因而，一般情况下，训练批次大小越小，则表明一个批次样本中的随机性越大，最终会导致损失函数越不易收敛。训练批次大小越

大，则越能够表征全体数据的特征，梯度下降的方向相对会越准确，损失函数的收敛速度也相应会越快。但同时由于缺乏随机性，很容易导致梯度一直沿某个单一方向下降，模型陷入局部最优；而且当训练批次大小增大到一定程度时，也会导致一个批次产生的权重更新停止变化或变化很小。因此，需要通过不断调整找到一个最恰当的训练批次大小，使得模型能够收敛速度最快并且效果最好。学习速率会影响模型能以多快的速度收敛至局部最小值，其值越小，表明沿梯度下降的步长就越小，进而获得最小值需要迭代的次数就越多，损失函数收敛所耗费的时间就越长，值越大，表明沿梯度下降的步长越大，会因收敛速度过快导致模型错过最佳收敛点，模型效果不稳定。因此，也需要选择合适的学习速率，使得模型训练的时间最少且稳定性最好。训练周期数决定了模型训练的总步数，当模型使用预热学习和线性学习率衰减方法进行训练优化时，模型训练的总步数又决定了预热学习的持续时间，需要考虑调整训练周期数，使其不至于太大导致前期学习率低且基本不优化，也不至于太小，使得训练结束太早，模型难以收敛和达不到预期效果等。

（2）Step2：配置GPU或TPU。BERT不支持在CPU上进行训练，谷歌训练BERT预训练模型使用的是TPU，所以原始模型中有很多TPU相关的配置，TPU通常比普通GPU具有更多的RAM，使用TPU训练效率会更高一些。一般情况下，根据具体任务和可用资源的情况，还可以通过减少训练批次大小或最大序列长度以适应实际硬件环境，但这会导致模型性能的下降。若在多个GPU上训练，还需初始化分布式后端以负责同步相应GPU节点或根据GPU的编号，选择只使用某些GPU。因此在创建和加载模型前，需要首先有根据相关参数设置检查和判断CUDA、GPU等设备是否可用和是否启用的条件，然后再定义和调用相应设备，并确保数据和模型在训练开始前已被移动到正确的相同的设备。

（3）Step3：创建模型，一个基于BERT的用于文本序列分类的方法模块。该模块的主要功能是通过配置文件和分类标签列表加载BERT预训练模型；根据配置文件中参数设置模型的dropout比率，以避免模型在训练中产生过拟合问题；在BERT模型的最后输出层上增加一个线性层作为分类器用于分类训练。分类器最常用的最基础的损失函数一般为交叉熵，能够表征真实样本类别和预测类别概率分布之间的距离，距离越小，交叉熵越小，表明两者之间越接近。但交叉熵用于多分类任务时，通常计算的是每一个样本对应于每个分类标签的归一化结果，即需要假设和保持每个样本属于多个分类标签的概率总和为1，最后哪个结果值大，则将其判定到哪一类。但实际样本中，往往每个样本的多个分类标签之间可能是相互独立的，没有明显关系，所以概率之和不一定为1。

基于上述考虑，本文主要使二进制交叉熵来取代原模型的交叉熵函数，可以在计算损失时为每一个分类标签分配独立的概率，使得模型的适配性和数值计算的稳定性更好。二进制交叉熵的计算方法如公式（1）所示。

$$\text{loss(ouput,target)} = -\frac{1}{n}\sum_i \text{weight}i(\text{target}_i * \log(\text{sigmoid}(\text{output}_i)) + (1-\text{target}_i) * \log(1-\text{sigmoid}(\text{output}_i))) \quad (1)$$

其中，output表示原BERT模型输出，实质是对其再增加一个激活函数sigmoid运算，用于将原输出的实数向量映射到(0,1)的区间，在特征比较复杂或相差不是很大时效果较好。当损失计算不进行输出缩放（即降维）时，将与原输出向量具有相同的维数。target表示预测的目标类别。weight表示可以为每一批次元素的损失手动调整权重，给定值必须为等同于批次大小的张量，默认情况下，也可以不给定，所有批次的权重都为1。

（4）Step4：设计优化器，即模型的训练优化算法。深度神经网络的优化算法通常是由一个反向传播算法与一个梯度下降算法组成。一般的浅层神经网络通过梯度下降算法即能够使网络参数不断收敛至全局或局部最小值，但由于深度神经网络层数太多，还必须通过反向传播算法把误差逐层地从输出反向传播到输入以实现逐层地更新网络参数。目前主流的深度学习优化算法有Adam算法、动量算法、RMSprop算法、梯度下降算法、同步聚合梯度算法等。不同优化算法的特点不同：如梯度下降算法是最原始最基础的算法，是一次性将所有训练集都载入并计算全部的梯度，由于梯度方向是函数值变大的最快的方向，因此负梯度方向则是函数值变小的最快的方向，因此只要沿着梯度相反的方向逐步迭代更新权重，算法函数就能够收敛到最小值，但缺点是计算量太大，普通的GPU难以承载算法运行。如同步聚合梯度算法主要应用于异步训练场景，常规地对于N个训练集的异步训练，梯度将独立地应用到变量N次，该算法通过从所有训练集副本中收集梯度并进行平均，然后一次性地将其应用到变量中，从而实现了梯度的快速更新并且避免了梯度过时。如动量算法主要是从训练集中采集一定数目小批量样本，采用指数加权平均方法来计算梯度，比标准梯度下降算法具有更快的收敛速度。如RMSprop算法也是从训练集中采集一定数目小批量样本，主要采用归一化方法来计算梯度，并增加了衰减率指标，可以通过调整不同维度上的步长加快收敛速度，但由于梯度衰减问题，会导致后期训练速度变慢。本文主要基于Adam算法实现了一个对BERT模型进行权重衰减修正的训练优化方法BERTAdam。Adam算法实质上是动量算法和RMSprop算法的结合，集成了两者的优点，相当于先对原始梯度进行指数加权平均，接着进行归一化处理，最后再更新梯度，可以有效避免一开始梯度小导致的学习步长较大，后期收敛效果不好的问题。Adam的权重w迭代更新的计算方法如公式（2）所示：

$$\begin{cases} v = b_1 v + (1-b_1)dw \\ s = b_2 s + (1-b_2)dw^2 \\ w = w - \alpha \dfrac{v}{s+\varepsilon} \end{cases} \quad (2)$$

其中，b1和b2表示Adam进行矩估计的指数衰减速率，e表示学习步长。一般情况下，b1 = 0.9，b2 = 0.999，e = 1e-6。

BERTAdam除了具有Adam算法的固有参数b1、b2、e

外，主要参数还有：学习速率、预热学习率、训练总步数、权重衰减速率等。其中，训练总步数=训练集的实例总数/训练的批次大小/(梯度累积的步数*训练周期数)。预热学习的意义是：由于刚开始训练时，模型的权重是随机初始化的，此时若选择一个较大的学习率，可能导致模型的不稳定，选择预热学习的方式，可以使得开始训练的几个周期或者一些步数内学习率较小，在预热的小学习率下模型就可以慢慢趋于稳定，等到模型相对稳定后再选择预先设置的全局学习速率进行训练，可以使得模型收敛速度变得更快，模型效果更好。

(5)Step5：加载训练数据。首先，调用3.2.3小节的数据处理与特征表示方法，读取训练数据，形成特征表示，并进一步转换为深度神经网络可读的张量及张量集。接着，设计采样器，即模型的训练采样方法。目前主流的采样方法有随机采样、顺序采样、分布式采样、权重随机采样等。其中随机采样是最常使用的采样方法，比较简单，是在全部样本中进行随机采样，采样的范围比较广。顺序采样是指按样本元素排列的固定顺序进行采样，分布式采样主要用于分布式深度学习时进行采样，权重随机采样主要是根据样本类别的权重来确定从哪个样本类别中进行采样，主要用于样本类别比例不均衡时进行重采样的场景。本文主要使用随机采样方法来设计采样器。最后，将训练张量集、采样器、训练批次大小等作为参数，传入模型的数据加载器，实现训练数据加载。

(6)Step6：模型训练与评估。训练的第一步是根据预设的训练批次大小，从已加载的训练数据中生成一个批次的数据，进行训练与计算损失等，然后迭代该过程，直至全部训练周期结束。每一轮训练中，如果使用的GPU节点大于1，则当前损失等于多个GPU上的损失均值；如果同时使用了32位和16位浮点精度的混合精度训练，则权重、梯度、激活函数等一般都会被保存为16位浮点精度的形式，在损失缩放大于1的时候，则当前损失=既有损失*损失范围；当梯度累积的步数>1时，则当前损失=即有损失/梯度累积的步数。然后继续进行梯度的反向更新。同时，每一轮训练结束后会在验证集上进行评估验证，主要是计算相应的F1值，如果F1值大于此前最高分则保存模型参数，否则设置一个标志位flags将其加1。如果flags大于某个预设常数N，即连续N轮内，模型的性能都没能够继续得到优化，则停止训练过程。N一般设置为8或根据实际情况来增加或减少。最终训练结果除了得到分类模型外，还可以计算出当前模型的评估精度、评估损失值、训练与优化的全局步数、平均损失值等。

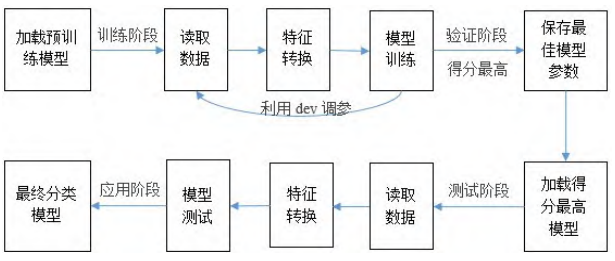

图3 基于BERT的分类关系自动识别模型的训练与优化流程

(7)Step7：模型测试与应用。测试过程与训练阶段相似，需要加载训练阶段的得分最高模型，读取测试数据进行数据处理与特征表示，转换为特征向量，输入模型，最后根据测试结果对模型进行评分。应用过程，则是直接加载最终的分类模型，对无标签数据进行预测分类，获得其分类标识。

4 领域本体分类关系自动识别的实验应用研究

4.1 实验数据集的获取与预处理

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分类标注语料集，需自行构造语料集。首先，邀请了资源环境学科领域的老师参与制定了资源环境学科领域的初始化本体分类体系，用以作为领域本体分类关系识别的基础支撑框架。其次，为了保证实验所用的领域语料具有一定的权威性、专业性，主要采集了以核心期刊/会议论文为主的学术文献类语料。主要通过Web of Science数据库，选择中国科学引文数据库CSCD，以该分类体系为基础分别构造检索式进行高级检索，时间跨度为2010-2019，基于Web Services API编写接口调用和解析程序对学术论文相关的元数据进行自动采集和抽取。如以“农业科学”为例，构造的主题检索式为“SU= AGRICULTURE”。同时，处于对样本的均衡性考虑，应保证每个分类下的数据总量都相差不大。最后，将获取到的学术文献的文摘及对应的中文类目名称以3.2.2节所述格式生成3个TSV文件，分别作为训练集(train.tsv)、验证集(dev.tsv)和测试集(test.tsv)，并保证每个类别下3个集中数据比例大概为8:1:1。删除一些出现乱码的文本序列，最终得到的有效训练集的文本序列数为304204个，验证集的文本序列数为38855个，测试集的文本序列数为39596个。将3个tsv文档存放到服务器指定路径$CORPUS_DIR/resource待用。

4.2 实验方法、工具与关键过程

实验服务器为ThinkStation P720塔式工作站，Linux操作系统，64G内存，NVIDIA TITAN RTX 24G GPU。实验的关键过程如下：

首先，BERT预训练模型的获取与准备。如3.2.1小节所述，目前谷歌在Github上已开源了一个BERT中文预训练模型base版。但该模型自最初发布后再没有更新过，训练使用的数据量、训练批次都相对比较小，且模型训练使用的中文语料是以字为粒度进行切分，并没有考虑到中文分词的问题。调研发现，随着BERT得到高度关注，国内外实力比较雄厚的相关研究团队已开始基于BERT及BERT原始的中文预训练模型，通过改进训练任务和数据生成方式、语言模型遮蔽方式、训练时间更长久、训练批次更大、训练语料更多、训练优化器更多等，陆续训练出了一批覆盖领域更多样化，性能更强大的中文预训练模型。如清华大学人工智能研究

院公开的百度百科BERT[20]。如哈尔滨工业大学讯飞联合实验室公开的BERT-wwm、BERT-wwm-ext等[21]。近期，美国Facebook AI和华盛顿大学联合发布了RoBERTa-zh-Large、Roberta_l24_zh_base等[22]。基于上述研究，本文选择目前最新也是最大的RoBERTa-zh-Large作为预训练模型，将其预先下载到服务器指定目录$BERT_DIR待用。其次，从Github上下载BERT源码，导入PyCharm，按3.2.3小节所述，加入对实验语料数据进行预处理、特征表示和转换的自定义程序，按3.2.4小节所述，准备模型的超参数，配置GPU，创建模型，设计优化器，设计采样器等，为微调BERT模型做准备。接着，通过命令和超参数设置，运行BERT分类脚本，加载试验语料和RoBERTa-zh-Large，执行微调，对模型进行训练与优化。最佳模型参数将输出和保存在$MODEL_DIR/bert_output目录。所用命令参数如图4所示。

```
python run_classifier.py
--task_name=RESOURCE
--do_train=true
--do_eval=true
--data_dir=$CORPUS_DIR/resource
--vocab_file=$BERT_DIR/roeberta_zh_L-24_H-1024_A-16/vocab.txt
--bert_config_file=$BERT_DIR/roeberta_zh_L-24_H-1024_A-16/bert_config_large.json
--init_checkpoint=$BERT_DIR/roeberta_zh_L-24_H-1024_A-16/roberta_zh_large_model.ckpt
--max_seq_length=128
--train_batch_size=32
--learning_rate=2e-5
--num_train_epochs=6.0
--output_dir=$MODEL_DIR/bert_output
```

图4 基于BERT的领域本体分类模型训练示例

4.3 实验模型的评估与结果分析

BERT本身已提供了对分类模型的评估程序，只要在上述训练优化过程中通过参数--do_eval=true开启评估，则会自动基于验证集对模型进行评估，并输出AUC、准确率、损失值等。其中AUC(Area Under Curve，曲线下面积)，是目前业内用来评估分类模型是否有效的一个最常用指标，AUC值越接近于1，通常表明分类模型的真实性越好、应用价值越高，小于等于0.5时，则表明真实性比较差，几乎无应用价值。本文也使用了相同语料，在先前的经典深度学习分类模型上进行了一系列实验，以便进行比对，总体实验结果如表1所示：

表1 资源环境领域本体分类关系自动识别模型的实验结果

分类模型	预训练模型	AUC	准确率	损失值
TextCNN	Word2Vec	0.9925	0.3030	4.0673
TextRCNN	Word2Vec	0.9953	0.4271	3.0106
FastText	------	0.9946	0.3709	3.0243
TextRNN	Word2Vec	0.9954	0.4516	2.8617
DPCNN	Word2Vec	0.9967	0.7956	0.0676
DRNN	Word2Vec	0.9969	0.8174	0.0664
AttnConvnet	Word2Vec	0.9978	0.8456	0.0514
BERT-Based	BERT-wwm-ext	0.9987	0.9961	0.0181
BERT-Based	RoBERTa-wwm-large	0.9995	0.9990	0.0094

需要特别强调的是，表1中涉及到基于BERT的分类模型，本文通常将其训练周期设置为3/6，训练批次为8/16/32，学习速率为2e-5/3e-5/5e-5，最大序列长度为128/256。涉及到其他深度学习模型，需要预训练模型的都使用Word2Vec，其他训练方法和参数取值范围不一样，皆借鉴前人研究，将其取值调整到相应算法支持的最大值或最优值。最终表中列出的精度和损失值为每个分类模型在验证集上通过多次调参实验后取的最优值。实验结果表明，同等条件下，其一，基于BERT的分类模型的分类精度和效果要远远高于其他传统分类模型，可见BERT是值得被改进和扩展应用的。其二，对于基于BERT的分类模型来说，训练批次越大，训练周期越长，单次读取语料的最大序列长度越长，分类模型的精度会得到逐步提升。同时，本文还发现最终分类模型精度是否有提高与当前BERT使用的预训练模型的数据量大小关系并不大，因为同等条件下使用大号版RoBERTa-wwm-large微调得到的结果与使用基础版BERT-wwm-ext得到的结果差别并不明显，精度也没有显著提升。因此，可见也不是一味盲目增大所有参数就一定能获得最优分类结果，应在实际应用时根据具体资源情况来合理调整参数。

此外，本文还做了其他一些对比实验：如使用搜狐语料库提供的新闻分类语料集取代自己构造的实验语料进行了训练与评估，基于BERT的分类模型模型精度为0.928，损失值为0.222。如使用著名机器学习竞赛平台Kaggle提供的基于维基百科评论的毒性评论语料集也同样进行了训练与评估，基于BERT的分类模型精度为0.987，损失值为0.076。对比实验结果表明，其一，基于BERT的分类模型具有极大的通用性，在其他领域仍能取得较高的分类精度。其二，语料集的质量对分类模型的精度影响非常大，只有大数据量、高质量的语料集才能获得最好的深度学习分类效果，也侧面表明本文通过机器程序自动获取、构建和预处理生成的多标签分类标注语料是相对准确的，技术路线是可行的，未来这套自动化的语料生成方法流程放在其他领域也将同样适用。

总的来说，基于BERT的深度学习机制改进和构建的领域本体分类关系识别模型，证明了从海量领域文本数据中自动学习和获取到领域本体分类关系的方法和路线是十分可行的，能够极大地减少人工的工作量和时间成本，具有极好地通用性和可移植性，并且与传统的分类方法相比分类精度和效果有了跨越式提升。未来，要进一步提高模型的精度，除了需要继续调整和优化模型相关的参数及方法外，可能还需要考虑建立对语料集质量，主要是分类标注的准确性进行严格审查和修正的机制，以尽力提升有效样本集的数量，从数据源头上确保模型训练的可靠性和有效性。

参考文献

1 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.

2 王思丽,祝忠明,刘　巍,等.基于深度学习的领域本体概念自动获取方法研究[J].情报理论与实践, 2020,43(3):145-152.

3 OpenGALEN[EB/OL].https://en.wikipedia.org/wik

i/OpenGALEN,2019-12-04.
4 王庆林,张九天.气候变化领域本体手册[M].北京:北京理工大学出版社,2012.
5 科技知识组织体系共享服务系统(STKOS)[EB/OL]. http://stkos.las.ac.cn/stkosservice/user/welcome.htm, 2019-11-26.
6 刘大伟. 基于WordNet本体库的文本分类方法[D]. 北京:北京交通大学, 2008.
7 黄浩军. 一种基于维基百科的文本分类算法[D]. 北京:北京大学, 2014.
8 Jun SY, Aliyeva D, Ji JM, et al. Utilizing Probase in Open Directory Project-based Text Classification [EB/OL].https://arxiv.org/abs/1805.04992,2018-05-14.
9 Kim Y. Convolutional Neural Networks for Sentence Classification[EB/OL].https://arxiv.org/abs/1408.5882,2014-07-03.
10 Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks forEfficient Text Classification[EB/OL]. https://arxiv.org/abs/1607.01759,2016-08-09.
11 Liu PF, Qiu XP, Huang XJ. Recurrent Neural Network for Text Classification with Multi-Task Learning[EB/OL]. https://arxiv.org/abs/1605.05101,2016-05-17.
12 Pappas N, Andrei PB. Multilingual Hierarchical Attention Networks for Document Classification[EB/OL].https://arxiv.org/abs/1707.00896,2017-09-05.
13 Kim Y, Lee H, Jung K. AttnConvnet at SemEval-2018 Task 1: Attention-based Convolutional Neural Networks for Multi-label Emotion Classification[EB/OL].http://sciencewise.info/articles/1804.00831/,2018-01-01.
14 Devlin J, Chang MW, Lee K, et al. BERT: pre-training ofdeep bidirectional transformers for language understanding [EB/OL]. https://arxiv.org/abs/1810.04805,2019-05-24.
15 Yang ZL, Dai ZH, Yang YM, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding[EB/OL]. https://arxiv.org/abs/1906.08237,2019-06-19.
16 Liu LQ, Mu FN, Li PY, et al. NeuralClassifier: An Open-source Neural Hierarchical Multi-label Text Classification Toolkit[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, 2019: 87-92.
17 BERT[EB/OL].https://github.com/google-research/bert/, 2019-11-26.
18 王思丽,祝忠明,刘　巍,等.领域本体学习语料的自动获取与预处理方法研究[J].图书馆学研究,2019,(20):54-64.
19 GLUE data[EB/OL].https://gluebenchmark.com/tasks,2019-11-26.
20 OpenCLaP [EB/OL]. https://github.com/thunlp/OpenCLaP,2019-11-26.
21 Pre-Training with Whole Word Masking for Chinese BERT[EB/OL].https://github.com/ymcui/Chinese-BERT-wwm,2019-11-26.
22 RoBERTa for Chinese[EB/OL].https://github.com/brightmart/roberta_zh,,2019-11-26..

(责任编辑:毛秀梅)