NLP技术及其在找矿预测中的应用研究

李　科

湖北省地质局第二地质大队　湖北恩施　445000

【摘　要】找矿预测在矿山开发中有着十分重要的作用，目前存在大量的地质资料，具有应用在找矿预测中的作用，可以采用智能算法对文本内容进行分析，确定矿脉的位置。本文探索一种自然语言处理（NLP）技术进行找矿预测的技术。利用Word2dev和BERT建立模型，使用收集到的地质论文、说明文献、书籍内容进行模型的训练，之后进行地质图向量化处理并使用之前的模型进行找矿预测。通过和之前的结果进行对比，使用NLP技术的预测结果和之前人工预测结果相吻合。说明使用NLP技术可以应用在找矿工作中，并且能够有效降低人员工作的强度，提升找矿工作效率。

【关键词】自然语言处理；人工智能；找矿预测

随着地质调查、找矿工作的实践数量的增加，实现了大量高质量、多尺度、地学相关数据的积累，给找矿勘探工作提供了大量的线索和数据。但是大量的报告、文档、文献等资料很少能够应用到找矿预测当中，使得目前文字资料所蕴含的信息还值得继续挖掘。使用自然语言处理（NLP）能够用自动的方式实现对自然语言的加工，可以满足特殊情况的需要。本文就用NLP技术结合神经网络探索找矿预测技术，从而充分利用文字信息，提升找矿工作的效率。

1　基于Word2vec的语料库构建

1.1　Word2vec模型

Word2vec模型使用神经网络特定参数当作词语的分布式特征，可以处理词语和语义之间的相似度，目前Word2vec模型包括三层，分别是输入层、投影层和输出层，CBOW模型会根据当前词w(t)的上下文预测当前词，使用skipgram模型会根据当前词进行上下文预测。

1.2　语料库构建

为了针对某矿区地质图向量化处理，收集了西秦岭、寨上、锁龙、卡琳型金矿等关键词的文献，查阅地区的地质报告、说明书，查找各类找矿预测理论书籍。对上述资料统一转为txt文本格式，以方便进行计算处理，之后统一存放，利用python中的os函数库对路径下的文件进行读写合并，获得生语料。处理过程中，删除语料中的乱码、空行、特殊字符等无法利用的文本，并使用Jieba函数库完成分词。

1.3　词向量获取

使用skipgram模型进行词嵌入，该模型输入词的独热码，可以获得某些词语的向量。通过训练能够获得词汇的编码向量，最终输出模型是概率分布，可以获得上下文取词窗口中出现某些词汇的概率。训练好网络模型后，不能直接进行上下文本预测任务，还需要词向量来进行权重参数变换。对于训练好的300维词向量，存在维度过高不便于观察的问题，因此需要通过降维工具将词汇的维度从300维降低到二维。

1.4　词向量应用

通过使用训练好的模型文件，能够得到词和词之间的相似度，通过使用Gensim能够从原始的非结构化文本中，无监督地学习到文本隐藏层的主题向量表达，使用similarity函数能够计算词向量之间语义相似性，还能够选取连续性规律的地质名进行语义相似度的计算。

1.5　简易问答系统搭建

通过使用向量和语义相似度计算，可以建立起简易的检索式问答系统，系统会读取文档内容，加载停用词，以及进行文档分词，并且将文档进行pickel序列化，并将语句储存成字典，根据语句中的词向量均值作为单句的向量标识。在用户提出问题时进行分词，将问题放入列表中，对问题各词向量使用句向量标识，并且对问题句向量和原文的向量进行比较，完成对问题结果的定位。

2　基于BERT模型的句向量应用

BERT是一种训练语言标识方法，能够获得文本数据，并从文本数据中获得语言特征，用户根据需要，可以使用该模型进行文本的微调，满足特定任务的需要。根据情况，可以进行文本分类、标注、特征提取、问答等工作。使用BERT能够进行深层次的学习，而且能够避免GPT等其他网络从左到右单向训练的局限，能够通过双向训练满足应用要求。

2.1　BERT模型微调

BERT训练的能耗比较大，为此需要结合地质领域的应用需求，进行BERT模型的微调。目前哈工大发布的BERT-wwm模

型，在结构上与原始BERT保持一致，却在处理中文文档特别是地质文档上表现出更高效率。该模型作为基础，经过进一步训练以适配地质专业语料，形成了地质学领域专用的BERT-GEO模型。通过对大量的地质文献、技术说明书和专业著作的深入学习，BERT-GEO模型有效地内化了地质学的语言模式和知识体系。这一过程不仅加深了模型对地质术语的理解，还提高了其在地质文本处理中的准确性和效率。尤其在进行地质图解释和矿产资源预测时，BERT-GEO表现出较原始BERT更优的性能。微调后的BERT-GEO模型不仅提高了能源利用率，还增强了模型在地质领域的应用价值，为找矿预测提供了一种更为精确和经济的技术途径。

2.2　句向量获取和可视化

BERT模型可以得到字词句式的不同粒度文本嵌入向量，使用相关词库可以对BERT映射出的向量做出规定，提升表示的清晰度。根据需要，将BERT句向量降维到100维，并进行后续计算。

3　地质图向量化和找矿预测

3.1　地质图网格化处理

针对某地地层进行网格化处理，将地层、岩体、岩株、断层破碎带等地质实体转换成统一的数字格式。具体而言，八个不同的地质图被合并为一个统一的MapGIS矢量文件，这一步骤是实现地质图的全面数字化的基础。同时，地质图的柱状图、图例和相关报告性文字也被整合，以确保所有相关地质信息的全面性和连续性。接着，通过Word2Vec技术提取地质数据的句向量，将地质文本数据转换为向量格式，为后续的深度学习模型提供输入数据的重要步骤。然后，利用BERT模型来优化句向量的质量，BERT模型的高效处理能力能够提取地质文本中更深层次的含义。接着BERT-GEO模型的应用对于地质图的向量化处理至关重要。模型专门针对地质数据设计，能够更加准确地处理和解释与地质相关的复杂信息。利用BERT-GEO模型，可以实现地质图的高精度向量化表示，这对于找矿预测的准确性具有重大意义。最后，利用matplotlib工具进行模型结果的可视化，不仅使得数据更加直观易懂，还有助于展现地质结构和预测矿产分布。这一步骤是将地质图的数字化处理和找矿预测有效结合的关键，提供了一种更直接和高效的方法来分析和解释地质数据。

3.2　物化资料准备

为了方便进行运算，以及保证图件的完整性，采用具有全面数据、工作程度　较高的四幅1:50000图幅作为目标区，通过整理资料进行人工智能找矿预测。使用收集到的物探资料最为存放航磁数据坐标点的表格，物化资料主要为水系地球化学不同元素的含量。数据中提供了包括银、砷、金、钴、铬、铜、镍、钼等十三中元素，使用mapgis点文件属性字段坐标点、地球化学元素含量进行表示，并将数据储存为数据表。针对含有坐标、地化元素的　数据表，使用规定插值法，设定横纵坐标范围，作为目标区域的地质图，以100米为网格间距，获取网格化的数据。通过分析，很多元素存在的一场分布都和眼球区域内的金矿点位置关系密切。

3.3　卷积神经网络

卷积神经网络能够进行图片分类，以及目标检测和识别上获得良好的效果，网络由卷积层、池化层、全连接层组成，卷积审计网络会在卷积层中提取数据的局部特征，之后利用池化层简化特征数量，最后全连接层合并获得分类结果。

3.3.1　网络构建和找矿预测

在本研究中，运用了卷积神经网络（CNN），来预测地质数据中的矿产分布。CNN的网络结构采用了三层卷积层加上池化层，紧随其后的是全连接层，然后是三层flatten层，最后是dropout层和softmax层。在这个体系中，每个卷积层通过提取图像特征的层级结构，逐步增强了模型对地质数据的理解能力。每个卷积层后的池化层有助于减少计算量，同时保持特征的重要部分，提高了模型的泛化能力。全连接层将学习到的高层特征进行汇总，为最终的分类决策提供基础。flatten层将卷积和池化层的多维输出展平成一维，以便全连接层可以处理。dropout层是一种正则化技术，随机地“丢弃”网络中的一些节点，从而减少过拟合，提高模型的泛化能力。最终，softmax层将全连接层的输出转换为概率分布，这表明了每个分类目标的可能性，本研究中即为有矿或无矿的预测。使用交叉熵损失函数进行优化，该函数衡量了预测概率分布与实际标签之间的差异，通过梯度下降算法不断调整网络权重来最小化损失值。

3.3.2　卷积层

卷积层包括若干特征图，能够直接对原始信号进行卷积操作，每层卷积层包括多个卷积单元，每个卷积单元的参数都会通过反向传播算法获得。卷积层需要通过局部连接和权值共享的方法获得视觉特征。在本研究中，所使用的卷积核大小为3×3，卷积核会以固定某个步长在特征图上滑动，并进行卷积操作。多次输入的信号会产生不同的卷积操作，每个卷积核都会有自身特征。

3.3.3　激活函数

卷积之后会引入激活函数，进行非线性映射，该模型的

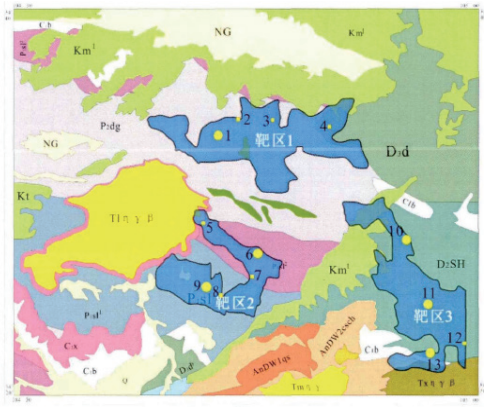

图1 NLP预测结果

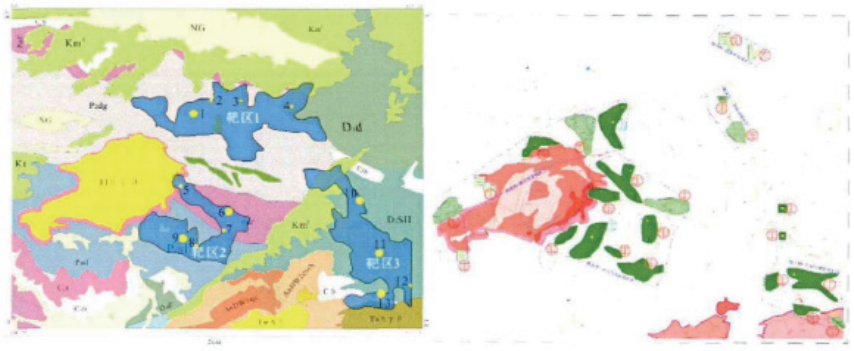

图2预测结果对比

激活函数为ReLU，能够保证训练结果的精确性，并加快计算速度，函数为：$f(x) = \max(0, x)$。

3.3.4 池化层

池化层能够从数据中过滤得到池化特征，并保留数据，池化的方法包括最大值池化和平均池化，经过池化层采样之后，输出特征图数量保持不变，但是能够减少计算复杂度。

3.4 数据处理和预测

3.4.1 数据集建立

以100m的网格尺度将目标区域分成370行、461列，提取区域内矿床的坐标，计算矿点所在网格的行列数，矿点所在网格为含矿网络，其他位置为位置网络。

3.4.2 数据增强

学习数据的获取是深度学习基础，基于深度学习找矿预测中找矿预测数量较多的情况下，一个区域仅存在少数几个已知矿点的情况，为了能够提升学习效率，采用通过窗口偏移生成学习数据的方法进行数据增强。通过观察矿床的地球化学特征，生成类似数据进行分析。并且，还在随机选取已知矿床点三倍网格单元作为未知区，增加训练数据的数量。

3.4.3 找矿靶区预测

训练集和验证集会根据已知数据6个近况位置计算它们所在网络，并使用数据增强的方法获得了3456个训练数据，其中70%用于对模型的训练，30%用于模型的验证。使用参数和数据集进行模型的200轮训练，结果显示经过50轮训练之后，模型已经能够获得比较稳定的结果，获得精度在98.1%左右。训练好的模型继续用于金矿矿区的找矿预测，预测区域占据目标区域总面积的17.54%。而且由于卷积神经网络窗口存在滑动机制，所以预测结果边界会略小于研究区域的实际范围。

3.4.4 预测结果分析

如图1所示，网络的预测结果揭示了三个具有潜在成矿能力的地区。在1号区域，观察到的不整合接触带显示出较高的含金量，这是由于地层特殊的地质构造导致金的积聚。该区域的岩石受到了强烈的破碎作用，这不仅为金矿的富集提供了必要的通道，而且还创造了空间，使其成为具有优秀成矿前景的地区。2号区域的预测结果显示存在褶皱断裂构造，这类构造通常与矿体的形成和分布有着密切的关联。褶皱断裂构造在地质历史上可能伴随着矿物质的运移和集中，这使得该区域也显示出了较好的成矿潜力。在3号区域，地质预测指出层间破碎带和断裂破碎带的发育。这些破碎带为矿物质的运移和沉积提供了通道，尤其是在有流体活动的情况下，破碎带可能成为矿物质沉积和富集的有利场所。因此，这一区域同样被认为是一个有潜力的成矿区。

通过使用前人预测结果和文本预测结果进行对比（如图2所示），1区域预测结果和金矿区重合；2区域和3区域范围内的成矿区分布也基本一致，证明使用NLP可以获得比较准确的预测结果。

结论：使用NLP技术开展找矿工作可以提升找矿工作的效率，优化找矿的模式，通过利用NLP技术和卷积神经网络结合，能提升找矿的效果。为了能够更广泛地使用NLP技术，需要加强语料库的构建，为开展预测提供更多、更全面的资料，也能推动进行网络的深层次训练，让网络可以提供更加准确的预测结果，满足地质工作需求。

参考文献：

[1]王晴,黄进,刘鑫等.成果地质资料知识图谱构建与可视化[J].计算机系统应用,2022,31(08):140-145.

[2]戴均豪.基于NLP技术的地质图向量化方法及其找矿预测应用[D].吉林大学,2021.