DOI：10. 12090/j. issn. 1006-6616. 2021. 27. 03. 035
文章编号：1006-6616（2021）03-0391-09

基于 BERT 的金矿地质实体关系抽取模型研究

黄徐胜[1,2]，朱月琴[3,4]，付立军[1,2,5]，刘雨江[1,2]，唐珂珂[1,2]，李　金[1,2]
HUANG Xusheng[1,2]，ZHU Yueqin[3,4]，FU Lijun[1,2,5]，LIU Yujiang[1,2]，TANG Keke[1,2]，LI Jin[1,2]

1. 中国科学院大学，北京 100049；
2. 中国科学院沈阳计算技术研究所，辽宁 沈阳 110168；
3. 自然资源部地质信息工程技术创新中心，北京 100037；
4. 中国地质调查局发展研究中心，北京 100037；
5. 山东大学大数据技术与认知智能实验室，山东 济南 250100
1. *University of Chinese Academy of Sciences*，*Beijing* 100049，*China*；
2. *Shenyang Institute of Computing Technology*，*Chinese Academy of Sciences*，*Shenyang* 110168，*Liaoning*，*China*；
3. *Technology Innovation Center of Geological Information*，*MNR*，*Beijing* 100037，*China*；
4. *Development Research Center of China Geological Survey*，*Beijing* 100037，*China*；
5. *Laboratory of Big Data and Artificial Intelligence Technology*，*Shandong University*，*Jinan* 250100，*Shandong*，*China*

HUANG X S，ZHU Y Q，FU L J，et al.，2021. Research on a geological entity relation extraction model for gold mine based on BERT［J］. Journal of Geomechanics，27（3）：391－399. DOI：10. 12090/j. issn. 1006-6616. 2021. 27. 03. 035

Abstract：Intelligent identification of entity relation is an important method and approach to improve literature mining and analysis，and knowledge extraction of gold mine. This study focuses on the core issues affecting current entity relation extraction of gold mine such as complex entity relation and less manual annotation information，and proposes a BERT（Bidirectional Encoder Representations from Transformer）remotely supervised relation extraction model. The accuracy of relation extraction is increased by optimizing and improving the modules related to geological data coding，geological classification and geological entity filtering. And the effectiveness of the model is verified by the entity relation extraction experiment of 290489 pieces of gold ore documents.

Key words：remote supervision；relation extraction；BERT；geological entity

摘　要：金矿实体关系的智能识别是提高金矿文献分析挖掘和知识提取的重要方法和途径。此次研究针对目前金矿实体关系抽取涉及到的核心问题，如金矿实体关系复杂、人工标注信息少等特点，提出了基于 BERT（Bidirectional Encoder Representations from Transformer）的远程监督关系抽取模型。并通过金矿地质数据编码、金矿分类和金矿地质实体过滤等模块的优化改进，提高了金矿地质实体关系抽取的准确率。最后通过对金矿文献数据的实体关系抽取实验，验证了该方法的有效性。

基金项目：国家自然科学基金项目（41872253）；国家重点研发计划项目（2018YFC1505501）；中国地质调查局地质调查项目（DD20190318）
This research is financially supported by the National Natural Science Foundation of China（Grant No. 41872253），National Key Research and Development Program（Grant No. 2018YFC1505501），and Geological Survey Project of China Geological Survey（Grant No. DD20190318）.
第一作者简介：黄徐胜（1994—），男，在读硕士，从事自然语言处理研究、人工智能在地质领域的应用。
E-mail：huangxusheng18@ mails. ucas. ac. cn
通讯作者：朱月琴（1987—），女，正高级工程师，从事地质大数据、人工智能在地学方面的研究及应用等工作。
E-mail：yueqinzhu@ 163. com
收稿日期：2020-11-20；**修回日期**：2021-01-10；**责任编辑**：范二平
引用格式：黄徐胜，朱月琴，付立军，等，2021. 基于 BERT 的金矿地质实体关系抽取模型研究［J］. 地质力学学报，27（3）：391-399. DOI：10. 12090/j. issn. 1006-6616. 2021. 27. 03. 035

关键词：远程监督；关系抽取；BERT；地质实体
中图分类号：P628.4 **文献标识码**：A

0 引言

地质文献是地质科研成果的规范化记录和表现形式，是探索研究地球科学的成果结晶和研究基础。随着新一代信息技术的研发及应用，各个部门积累了大量的地质文献数据。据统计，目前中国地质文献中心积累的地质文献数据约为9041万条，总量达102 T，属于典型的地质大数据（谭永杰等，2017；陈建平等，2017）。因此如何快速、有效地分析、挖掘这些海量的文献数据，发现潜在的地质知识价值，实现地质数据的“增值”，是地质信息化工作所面临的一项重要挑战。而基于地质文献的实体关系抽取研究的目的就是准确、高效识别并抽取出地质文献中的实体以及之间的关系，建立起实体间的知识体系结构，以便于人们快速发现和理解知识点之间的关联脉络，目前这已经成为地质大数据的一项研究热点。

金矿地质文献中蕴含了大量的金矿相关实体以及金矿地质实体之间的关联关系（薛玉山等，2020；张兵强等，2020；张康等，2020；汪青松等，2021），识别金矿实体及实体间的关系对于进一步挖掘金矿地质文献知识、提升金矿地质文献数据的分析挖掘以及促进金矿的进一步开采利用等方面有着积极而深远的意义。文中论述了一套基于远程监督关系的金矿地质实体关系抽取模型构建方法，尝试通过少量标注样本建立地质实体的关联关系的智能抽取方法，从而达到对金矿文献的快速分析挖掘及潜在知识发现的目的。

1 金矿地质实体关系抽取研究现状

关系抽取是信息抽取的核心内容，旨在提取文本中实体对的关系。在有监督关系抽取中，通常把关系抽取当作关系分类的问题来处理，但是模型经常面临缺乏训练标注数据的情况。为解决这个问题，Mintz et al.（2009）首次提出远程监督思想，使用知识库对齐目标文本的方法，构造远程监督数据集。Riedel et al.（2010）在此基础上提出“至少一次（at-least-once assumption）”假设，把远程监督关系抽取看作多实例学习（MIL）问题，把所有包含该实体对的句子整合成句袋，基于句袋进行分类。Hoffmann et al.（2010）提出多实例结合多标签的方法缓解错误标注问题。Zeng et al.（2015）在远程监督方法中使用CNN（convolutional neural networks）模型，提出了Piecewise CNN（PCNN）模型，采用max pooling的方法，保留细粒度的信息。在PCNN的基础上，Lin et al.（2016）融合注意力机制，对句袋中的每一个句子分配权重，权重的大小决定了该句在句袋中的比重，有效地缓解了数据集中的噪声问题。Feng et al.（2018）提出了强化学习的方式，该模型的实例选择器用于减轻噪声，使得模型更有效地训练数据，然后进行关系分类训练。蔡强等（2018）融合句子层次的注意力机制和词语层次的注意力机制提出多尺度注意力机制的方法，准确率在NYT-Freebase（NYT）数据集上达到了78%。

除了浅层模型之外，Huang and Wang（2017）、蔡强等（2019）提出一种基于深度残差神经网络的抽取方法，该方法利用残差神经网络获取特征。唐朝等（2020）在残差网络的基础上，融合BiGRU模型，在公开的数据集NYT上进行关系抽取，准确率比残差网络提升了2.9%。Bing et al.（2019）、钱小梅等（2020）采用DenseNet神经网络的抽取方法加深网络，解决神经网络中梯度消失的情况。

近年来，预训练模型成为关注的焦点。BERT（Jacob et al.，2019）是google提出的基于双向Transformer（Vaswani et al.，2017）的网络模型，该模型被证明能够有效地应用在大部分的自然语言处理任务中。Soares et al.（2019）提出了一个利用预先训练的BERT语言模型结合目标实体的信息来处理关系分类任务的模型。Alt et al.（2019）提出将Generative Pre-trained Transformer（GPT）模型应用在远程监督关系抽取中，该模型被证明可以有效地捕获文本的语义和语法特征。上述模型做出了很大的贡献，但是存在以下问题：①没有解决关系方向问题；②没有学习复杂的数据特征。

由于地质实体间关系复杂、类型多，采用智

能建模方法实现地质实体的关系自动识别难度大。因此目前基于地质文本信息的抽取研究主要集中在地质实体的抽取及可视化表达等方面，如：Zhu et al.（2017）提出的地质知识图谱构建框架及探索；张雪英等（2018）采用 DBN 模型实现了对地质实体信息的初步识别；而地质实体间关系抽取还停留在初步探索阶段，如：朱月琴等（2017）在地质数据语义模型中提出地质文本表达中的 6 种地质语义关系；吕鹏飞等（2017）采用统计语言模型和基于规则的方式提取三元组集合等。因此文章在前期研究的基础上，构建了基于远程监督的金矿地质实体关系抽取模型，并通过金矿地质文献的少量人工标注，探索并实现了金矿地质实体关系的智能化抽取。

2　远程监督关系抽取

关系抽取是知识图谱补全的重要环节，在自然语言处理领域具有重要的地位，如智能问答（Yih et al.，2015）、语义搜索（朱月琴等，2017）等。使用有监督的方法进行关系抽取需要大量的语料，这些语料完全依赖于人工的标注。然而，人工标注的方法只能构建少量的数据集。同时针对特殊领域的关系抽取，由于对标注人员的专业知识有一定要求，因此标注进展非常缓慢。远程监督的关系抽取方法（Mintz et al.，2009）可以使用金矿知识库与金矿地质文献对齐的方法自动标注数据集。该方法做出如下假设："如果金矿地质文献中的两个实体在对应的金矿知识库中存在着某种关系，则认为两个金矿实体在所有含有这两个实体的句子中都有这样的关系"。

如图 1 前两个句子所示：如果知识库中存在实体关系三元组（焦家式金矿、类型、破碎带蚀变岩型金矿），那么包含"焦家式金矿"和"破碎带蚀变岩型金矿"这个实体对的所有句子都会存在"类型"关系。该方法解决了缺乏地质领域数据集的问题。然而，使用远程监督关系抽取模型时，存在以下缺点。

（1）远程监督的假设过强，会存在标注错误的问题。如图 1 第三句所示：句子中存在"焦家式金矿"和"破碎带蚀变岩型金矿"实体对，"破碎带蚀变岩型金矿"并不是"焦家式金矿"的"类型"，但是仍然会以"类型"的关系存在于数据库中。这种启发式对齐知识库的方法，使得数据集存在错误标签问题。

（2）没有解决实体关系的方向问题。现有的关系抽取方法把关系抽取问题按照关系分类的方式处理，并不能很好地识别实体关系方向问题。除此之外，知识库和文本中的实体关系顺序存在不一致的可能。

（3）识别关系类别的难易程度不同。现有模型在训练过程中无法区分关系类别的易训练程度，导致模型不能有效地训练复杂的实体关系。

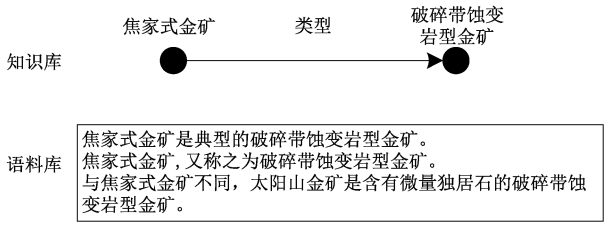

图 1　远程监督框架结构

Fig. 1　Framework of the remote supervision

3　基于远程监督的金矿地质实体关系抽取模型构建

根据金矿地质数据量大以及文献标注较少等特征，文中引入了远程监督的思想。针对金矿地质文献分析角度单一性和复杂性的数据特征，文中定义了金矿地质文献的实体和关系的类型，提出了结合地质领域特征的关系抽取模型。如图 2 所示，模型一共包括四个模块：①金矿地质数据编码模块；②基于 BERT 的金矿地质特征提取模块；③金矿地质分类模块；④金矿地质实体的过滤模块。

3.1　金矿地质数据编码模块

对于远程监督关系抽取，Zeng et al.（2015）曾通过分片卷积神经网络的方法，获取句子的结构信息，提高特征提取能力。文中结合金矿地质数据集中的知识，采用知识库顺序的实体编码方法。通过将一个无向关系分为两个有向关系的方式确定金矿地质文献中的关系方向。将每个关系 $r \in R$（R 为关系集合）分为两个关系类，即：$r(e_1, e_2)$ 和 $r(e_2, e_1)$，其中 e_1 和 e_2 表示两个实体。在实体前后加入特殊的标签"#"和"$"，"#"表示头节点的边界，"$"表示尾实体边界，实体根据知识库中的顺序进行标记，如图 2 所示。

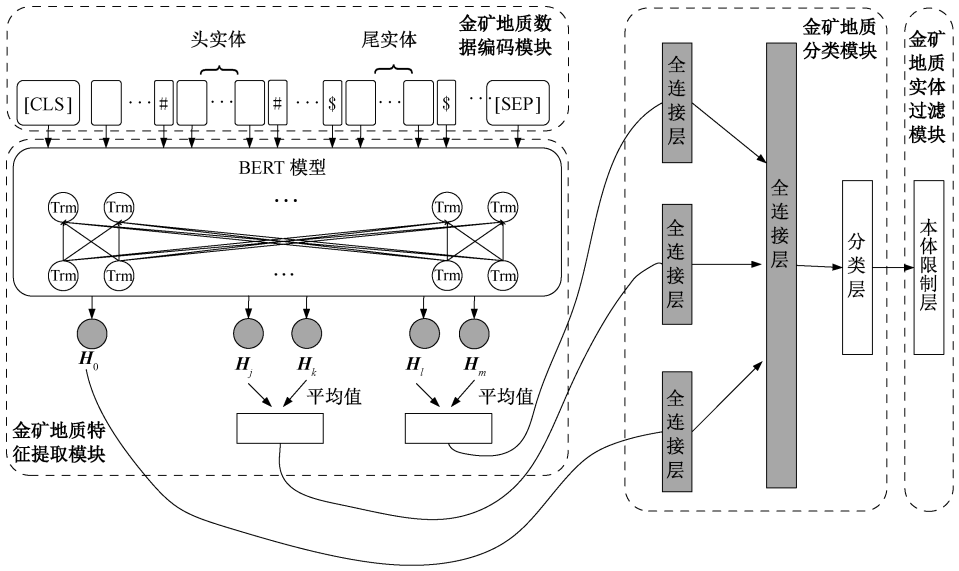

图 2 远程监督关系抽取模型

Fig. 2 Remotely supervised relation extraction model

3.2 基于 BERT 的金矿地质特征提取模块

BERT（Devlin et al.，2019）是 google 提出的基于双向 Transformer（Vaswani et al.，2017）的网络模型，在句子之间用［SEP］作为分隔符号，在每个序列的开始，添加一个特殊的字符［CLS］，用于存储该序列的语义信息。金矿的实体关系分类就可以利用［CLS］的输出进行预测。

模型的输入表示由词嵌入、句子嵌入和位置嵌入构成。给定一个包含头实体和尾实体的句子 s，$\boldsymbol{H}_j$ 和 $\boldsymbol{H}_k$ 向量是头实体经过 BERT 模型的状态向量，$\boldsymbol{H}_l$ 和 $\boldsymbol{H}_m$ 向量是尾实体经过 BERT 模型的状态向量。$\boldsymbol{H}_0$ 向量是［CLS］经过 BERT 的状态向量。经过平均操作、激活函数以及全连接层，得到最终的头实体和尾实体输出，如公式 1 所示：

$$\begin{aligned}\boldsymbol{H}_{\text{head}} &= \boldsymbol{W}_{\text{head}}\left[\tanh\left(\frac{1}{k-i+1}\sum_{t=i}^{k}\boldsymbol{H}_t\right)\right]+b_{\text{head}}\\ \boldsymbol{H}_{\text{tail}} &= \boldsymbol{W}_{\text{tail}}\left[\tanh\left(\frac{1}{l-m+1}\sum_{t=m}^{l}\boldsymbol{H}_t\right)\right]+b_{\text{tail}}\\ \boldsymbol{H}'_0 &= \boldsymbol{W}_0(\tanh(\boldsymbol{H}_0))+b_0\end{aligned} \tag{1}$$

其中 tanh 在神经网络中是激活函数，用于增加神经网络的非线性程度；$\boldsymbol{W}_{\text{head}}$ 表示头实体的权重向量；$\boldsymbol{W}_{\text{tail}}$ 表示尾实体的权重向量；$\boldsymbol{W}_0$ 表示［CLS］的权重向量；b_{head}、b_{tail}、b_0 表示头实体、尾实体和［CLS］的偏置参数；i、k 表示头实体的开始位置与结束位置；m、l 表示尾实体的开始位置和结束位置；$\boldsymbol{H}_t$ 表示 $\boldsymbol{H}_j$、$\boldsymbol{H}_k$ 这样的状态向量；$\boldsymbol{H}'_0$、$\boldsymbol{H}_{\text{head}}$ 和 $\boldsymbol{H}_{\text{tail}}$ 表示［CLS］、头实体、尾实体经过第一层全连接层得到的向量。

连接 $\boldsymbol{H}'_0$、$\boldsymbol{H}_{\text{head}}$ 和 $\boldsymbol{H}_{\text{tail}}$ 经过第二层全连接层得到向量 $\boldsymbol{H}_{\text{final}}$，如公式 2 所示：

$$\boldsymbol{H}_{\text{final}} = \boldsymbol{W}_{\text{final}}[\text{concat}(\boldsymbol{H}'_0, \boldsymbol{H}_{\text{head}}, \boldsymbol{H}_{\text{tail}})] + b_{\text{final}} \tag{2}$$

其中 $\boldsymbol{W}_{\text{final}}$ 为连接 $\boldsymbol{H}'_0$、$\boldsymbol{H}_{\text{head}}$ 和 $\boldsymbol{H}_{\text{tail}}$ 后计算的权重；b_{final} 为连接 $\boldsymbol{H}'_0$、$\boldsymbol{H}_{\text{head}}$ 和 $\boldsymbol{H}_{\text{tail}}$ 后计算的偏置参数；concat 是连接 $\boldsymbol{H}'_0$、$\boldsymbol{H}_{\text{head}}$ 和 $\boldsymbol{H}_{\text{tail}}$ 的函数。

3.3 金矿地质分类模块

训练神经网络的目标是使正确类的概率最大化，一般是通过最小化交叉熵损失（cross-entropy loss）来实现的，如公式（3）所示：

$$\text{CE}(p, y) = \begin{cases} -\ln(p) & \text{if } y = 1 \\ -\ln(1-p) & \text{otherwise} \end{cases} \tag{3}$$

其中 CE 表示交叉损失函数；p 表示准确率；y 表示是否为真实的标签。当 y 为真实的标签的时候，则进行 $-\ln(p)$ 运算，否则进行 $-\ln(1-p)$ 运算。

由于最小化交叉熵损失函数不能区分关系训练的难易程度，因此通过引入 γ 超参数识别样本难易程度（Lin et al.，2017），如公式（4）所示：

$$\text{loss} = -\alpha_t(1-p_t)^{\gamma}\ln(p_t) \tag{4}$$

其中 loss 为神经网络中的损失函数，用于判断预测值与真实值的差距；α 和 γ 是可以调整的超参数，α 为权重，介于［0，1］之间，用于减轻样本太多对训练的影响；t 为样本的编号，γ 用于对损失函数的调节。当 $y=1$ 时，p_t 趋向 1，表示容易训练的正样本，损失函数的权重趋向 0；当 $y=$

0 时，p_i 趋向 0，表示极难训练的正样本，对损失函数的权重趋向 0。因此通过该损失函数可以提取复杂地质实体之间的特性，从而大大提高了该模型的准确率。

3.4　金矿地质实体类别过滤

金矿地质领域存在丰富的实体关系信息。实体类别与实体类别之间存在着某种特有的关系，现有的模型并没有融合这些特征，造成抽取的实体关系存在常识错误等问题，准确率大大降低。结合金矿地质文献的特征，构建关系的实体类别过滤层。假设 $T=(E_1, E_2, \cdots, E_n)$，其中 T 是本体的集合。

如图 3，e_1，$e_2 \in E_1$；e_3，$e_4 \in E_2$；e_5，$e_6 \in E_3$；其中 e_n 代表金矿地质实体，如：黄铁矿、方铅矿、石英等。E_n 是实体所属的本体，如：化合物、金矿、种类等，r 表示的是实体间的关系，E_1 与 E_2 之间的关系是 $r_1(E_1, E_2)$，E_2 和 E_3 之间不存在关系，E_1 与 E_3 之间的关系是 $r_2(E_1, E_3)$。当抽取的结果是 $r_1(e_1, e_3)$，满足 E_1 与 E_2 之间的关系，设为可信。如果抽取结果为 $r_2(e_3, e_5)$ 时，其本体 E_2 与 E_3 没有关系，则将关系可信度设为 0。当抽取的结果是 $r_2(e_1, e_3)$，而 E_1 与 E_2 之间并没有这种关系，则将 $r_2(e_1, e3)$ 可信度设为 0，并在候选关系中，选择合适的关系。

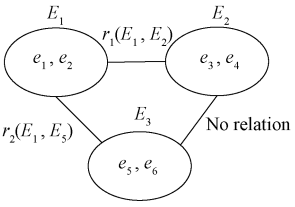

图 3　本体关系图

Fig. 3　Ontology diagram

4　模型的验证与分析

4.1　实验数据准备

首先，在 Riedel et al.（2010）数据集上评价文中建立的关系抽取模型。该 NYT 数据集由知识库 Freebase 和纽约时报语料库启发式对齐的方法构成，是远程监督领域标杆型的数据集。

同时，在地质领域数据集上进行模型检验。该数据来源于中国知网，选用 2000 至 2015 年 28740 篇与金矿相关的学术论文。经过数据预处理后，将人工标注的 4021 个金矿地质三元组与文献对齐，形成 290489 条数据集。从中挑选 50970 条有效数据，33761 条用于模型的训练，17209 条用于模型测试。文中概括归纳以金矿矿区、矿床、矿段和矿体为例的金矿地质实体一级关系为控矿因素、找矿标志以及属性特征等。其具体实体及关系见图 4 所示。

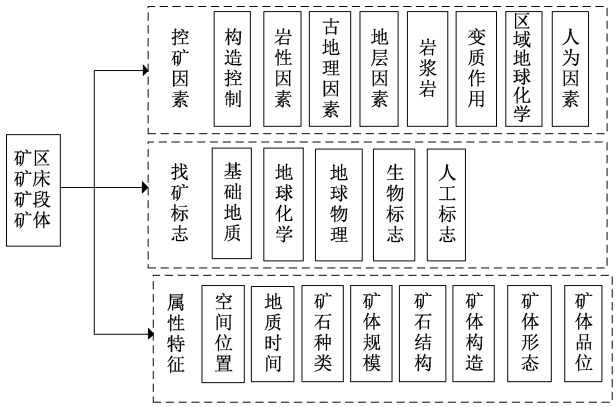

图 4　实体关系类别

Fig. 4　Categories of entity relation

4.2　实验过程与结果分析

文中将 Lin et al.（2016）提出的 PCNN+ATT（Piecewise CNN+attention）模型、Huang and Wang（2017）提出的残差神经网络（ResNet）模型和 Gao et al.（2017）提出的全连接神经网络（DenseNet）等模型作为基线模型，并对文章模型与基线模型的抽取效果展开了详细的解析。

（1）模型的实验参数设置

在这个实验中，文章模型实验的参数如表 1 所示。

表 1　实验参数

Table 1　Experiment parameters

参数名称	参数名称（英文）	符号	参数值
批大小	Batch_size	B	8
学习率	Adam Learning_rate	λ	2e−5
批次	Number of epoch	E	6
随机丢弃率	Droupout rate	P	0.1
最大句子长度	Max sentence length	ML	384

（2）模型的评价指标

模型的评价指标与其他远程监督关系抽取论文（Zeng et al.，2015；Lin et al.，2016；蔡强等，2018）指标类似，文中采用 $P@N$ 表示概率最大的前 N 个金矿地质实体关系预测正确的概率，分子表示预测成功的实体对的个数，N 需要手动设置，如：$N=100$，则表示前 100 个金矿地质实体对，如

公式（5）所示。PR（准确率-召回率）曲线图形成的面积可以用来评价模型的整体性能。

$$P@N = \frac{\text{前 } N \text{ 项被正确分类的数目}}{N} \times 100\% \quad (5)$$

（3）金矿地质实体关系抽取效果

为验证文中模型的关系抽取效果，分别在地质领域数据集和 NYT 通用数据集上进行对比实验，实验结果见表 2 和表 3。

表 2 各种模型在 NYT 数据集上的抽取效果

Table 2 Extraction effect of the models in NYT dataset

模型	接受者操作特征曲线下方面积大小（AUC）	Top N 项准确率（$P@N$/%）							平均准确率（Avg Prec/%）		
		100	200	300	500	1000	2000	5000	Top 300	Top 1000	Top 5000
DenseNet	0.34	81.0	69.5	68.7	61.4	51.6	39.5	22.4	73.1	66.4	56.3
ResNet	0.10	54.0	50.0	48.0	43.0	31.0	19.0	9.9	50.7	45.2	36.4
PCNN+ATT	0.32	74.0	67.5	64.3	59.8	48.7	37.2	22.3	68.6	62.9	53.4
文章模型	0.65	98.0	96.0	94.3	92.6	91.1	80.9	67.0	96.1	94.4	88.6

由表 2 可知，文中的方法在 Top 5000 时平均准确率达 88.6%，超过 PCNN+ATT 模型 35.2%。证明该模型能够一定程度降低数据集噪声，提高关系抽取的精准度。其次，文中的模型在 Top 300、Top 1000 和 Top 5000 上平均准确率分别为 96.1%、94.4%、88.6%，证明该模型有着更稳定的整体表现。综上所述，该方法在远程监督的任务上是可行的，可以用在地质领域的关系抽取中。

由表 3 可知，金矿地质实体关系抽取效果的整体趋势和 NYT 数据集一致。文中构建的模型在该领域数据集的关系抽取效果比 NYT 数据集好。原因是地质数据标签为 12 个，相比 NYT 数据集的 53 个标签，金矿地质实体关系分类相对容易。其次，地质数据的特征更加明显，特征提取更加方便。PCNN+ATT 模型虽然在 $P@100$ 时，平均准确率达到 99.0%，但是到了 $P@300$ 的时候，平均准确率急剧下降。相比之下，文章模型表现更加平稳，在 Top 300、Top 1000 和 Top 5000 时平均准确率分别为 100.0%、98.6%、93.1%。综上所述，文章模型在有向实体关系识别方面表现稳定。

表 3 各个方法在地质领域数据集上的抽取效果

Table 3 Extraction effect of the methods in geological dataset

模型	接受者操作特征曲线下方面积大小（AUC）	Top N 项准确率（$P@N$/%）							平均准确率（Avg Prec/%）		
		100	200	100	500	100	2000	100	Top300	Top 1000	Top5000
DenseNet	0.40	88.0	54.5	39.7	33.8	23.5	15.8	8.0	60.7	47.9	37.6
ResNet	0.34	70.0	52.5	40.3	30.6	23.6	15.3	7.8	54.3	43.4	34.3
PCNN+ATT	0.60	99.0	81.0	63.3	50.0	30.5	18.3	8.1	81.1	64.8	50.0
文章模型	0.75	100.0	100.0	99.3	98.4	98.6	96.1	93.1	99.8	99.3	97.9

由图 5 和图 6 可知，文中模型曲线为在 NYT 数据集和地质领域数据集上的面积分别为 0.65 和 0.75，在保证了准确率的同时，提升了召回率，因此该模型能够解决实体关系识别的长尾问题。

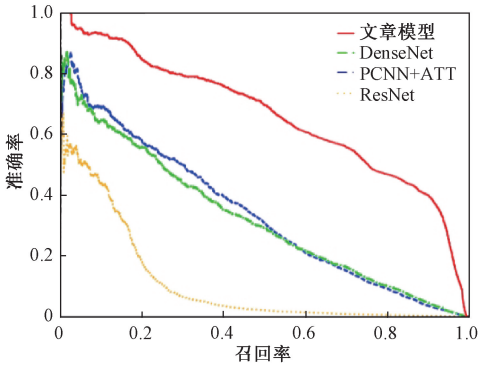

图 5 模型在 NYT 数据集上的 PR 图

Fig. 5 PR graph of each model in NYT dataset

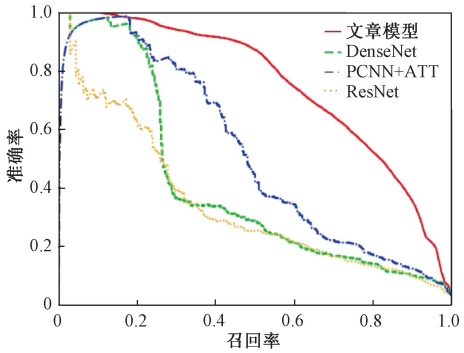

图 6 模型在地质数据集上的 PR 图

Fig. 6 PR graph of each model in geological dataset

使用文中模型对单篇文献数据（宋春明等，2014）进行关系抽取效果验证，部分抽取结果如图 7。由图 7 可知，模型对金矿地质实体和关系进行了准确的判断。

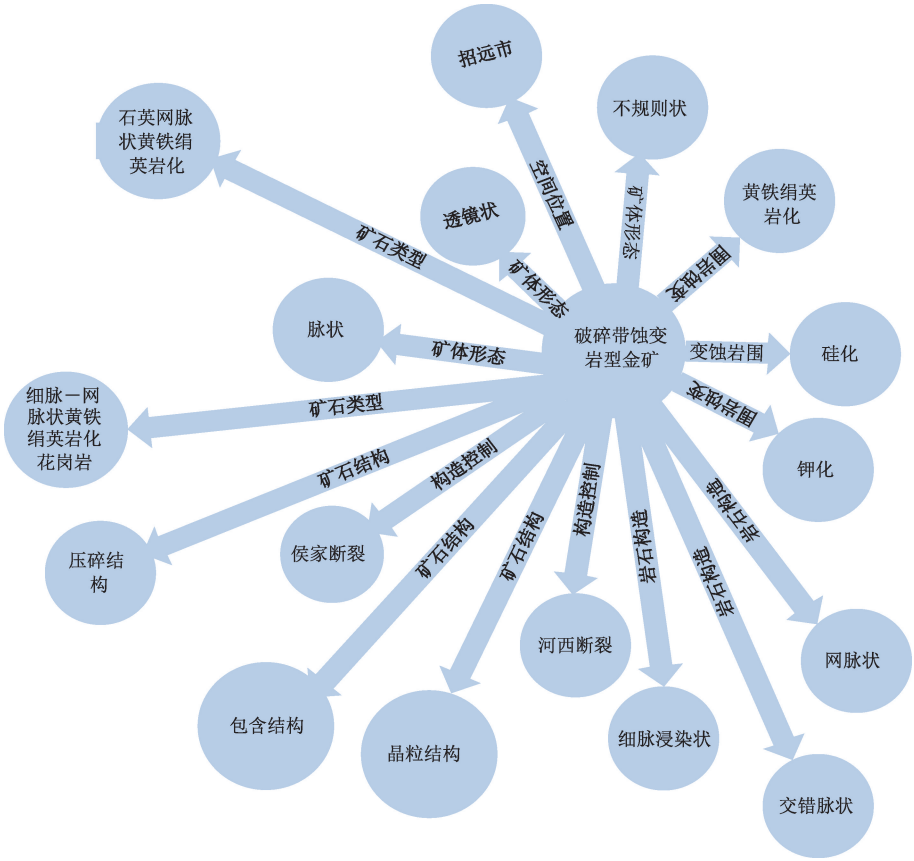

图 7　文章模型的抽取效果

Fig. 7　Extraction effect of BERT model

综上所述，文章的模型能够解决金矿地质实体关系抽取问题，在地质领域的数据集上有较高的准确率，长尾问题得到了解决。因此，该模型适用于金矿地质实体关系抽取。

5　结论

文章探讨了基于金矿文献的地质实体关系抽取方法。首次将远程监督关系抽取的思想引入金矿地质文献中，初步解决了目前由于金矿实体关系复杂、人工标注少而造成的金矿实体关系智能化抽取程度不高等问题；并利用远程监督的思想，构建了批量的金矿地质关系抽取实验数据集；同时在模型的构建及训练过程中，通过数据编码、分类模块、实体过滤以及限制输出等方法的改进，大大提升了金矿相关实体关系的抽取效果，实现了对金矿文献数据的实体关系抽取实验。实验结果验证了该方法的有效性。

将来可在进一步总结提炼金矿实体关系基础上，结合地质实体的背景信息，达到关系抽取效果的提升，从而为地质实体的智能识别、关系的抽取以及智能找矿等应用方面提供理论技术方法支撑。

References

ALT C, HÜBNER M, HENNIG L, 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction [C] //Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: Association for Computational Linguistics: 1388-1398.

BING X Y, SHEN L D, ZHENG L Y, 2019. A moderately deep convolutional neural network for relation extraction [C] // Proceedings of the 2019 11th international conference on machine learning and computing. New York, NY, USA: Association for Computing Machinery: 173-177.

CAI Q, HAO J Y, CAO J, et al., 2018. Multi-level attention mechanism based distant supervision for relation extraction [J]. Journal of Chinese Information Processing, 32 (1): 96-101. (in Chinese with English abstract)

CAI Q, LI J, HAO J Y, 2019. Distant supervision relation extraction based on focal loss and residual network [J]. Computer Engineering, 45 (12): 166-170. (in Chinese with English abstract)

CHEN J P, LI J, XIE S, et al., 2017. China geological big data research status [J]. Journal of Geology, 41 (3): 353-366. (in Chinese with English abstract)

DEVLIN J, CHANG M W, LEE K, et al., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding [C] // Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies. Minneapolis, Minnesota: Association for Computational Linguistics: 4171-4186.

FENG J, HUANG M L, ZHAO L, et al., 2018. Reinforcement learning for relation classification from noisy data [C] //Proceedings of the 32nd AAAI conference on artificial intelligence. Menlo Park, CA: AAAI: 5779-5786.

GAO H, LIU Z, VAN DER MAATEN L, et al., 2017. Densely connected convolutional networks [C] //Proceedings of the 2017 IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE: 4700-4708.

HOFFMANN R, ZHANG C L, LING X, et al., 2011. Knowledge-based weak supervision for information extraction of overlapping relations [C] //Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Portland, Oregon, USA: Association for Computational Linguistics: 541-550.

HUANG Y Y, WANG W Y, 2017. Deep residual learning for weakly-supervised relation extraction [C] //proceedings of the 2017 conference on empirical methods in natural language processing. Copenhagen, Denmark: Association for Computational Linguistics: 1803-1807.

LIN T Y, GOYAL P, GIRSHICK R, et al., 2017. Focal loss for dense object detection [C] //2017 IEEE international conference on computer vision (ICCV). Venice, Italy: IEEE: 2999-3007.

LIN Y K, SHEN S Q, LIU Z Y, et al., 2016. Neural relation extraction with selective attention over instances [C] //Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Berlin, Germany: Association for Computational Linguistics: 2124-2133.

LYU P F, WANG C N, ZHU Y Q, 2017. Study on geologic entity relation extraction method based on literature [J]. China Mining Magazine, 26 (10): 167-172. (in Chinese with English abstract)

MINTZ M, BILLS S, SNOW R, et al., 2009. Distant supervision for relation extraction without labeled data [C] //Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 2-Volume 2. Stroudsburg, PA: Association for Computational Linguistics: 1003-1011.

QIAN X M, LIU J Y, CHENG P S, 2020. Distant supervised relation extraction based on densely connected convolutional networks [J]. Computer Science, 47 (2): 157-162. (in Chinese with English abstract)

RIEDEL S, YAO L M, MCCALLUM A, 2010. Modeling relations and their mentions without labeled text [C] //Proceedings of the 2010 European conference on machine learning and knowledge discovery in databases. Berlin: Springer-Verlag: 148-163.

SOARES L B, FITZGERALD N, LING J, et al., 2019. Matching the blanks: distributional similarity for relation learning [C] // Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy: Association for Computational Linguistics: 2895-2905.

SONG M C, LI S Z, YI P H, et al., 2014. Classification and metallogenic theory of the Jiaojia-Style gold deposit in Jiaodong Peninsula, China [J]. Journal of Jilin University (Earth Science Edition), 44 (1): 87-104. (in Chinese with English abstract)

TAN Y J, WEN M, ZHU Y Q, et al., 2017. Research on the big data characteristics of geological data [J]. China Mining Magazine, 26 (9): 67-71, 84. (in Chinese with English abstract)

TANG C, NUO M H, HU Y, 2020. A hybrid model for relation extraction via ResNet & BiGRU [J]. Journal of Chinese Information Processing, 34 (2): 38-45. (in Chinese with English abstract)

VASWANI A, SHAZEER N, PARMAR N, et al., 2017. Attention is all you need [C] //Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.: 6000-6010.

WANG Q S, ZHANG J H, YOU T, et al., 2021. Study on the multiple-element exploration method of ore beds in wells and gold exploration experiment in the area with thick cover: Taken Wuhe area in Northeast Anhui as anexample [J]. Geology and Exploration, 57 (1): 136-145. (in Chinese with English abstract)

XUE Y S, WANG R T, WANG C, et al., 2020. Ore-controlling rules of fault structures in the Wangjiaping gold deposit in Shanyang County, Shaanxi Province [J]. Journal of Geomechanics, 26 (3): 391-404. (in Chinese with English abstract)

YIH W T, CHANG M W, HE X D, et al., 2015. Semantic parsing via staged query graph generation: question answering with knowledge base [C] //Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers). Beijing, China: Association for Computational Linguistics: 1321-1331.

ZENG D J, LIU K, CHEN Y B, et al., 2015. Distant supervision for relation extraction via piecewise convolutional neural networks [C] //Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics: 1753-1762.

ZHANG B Q, YANG Q H, ZHAO F Y, et al., 2020. The ore-bearing horizon and ore characteristics of gold deposits in the Emesishan basalt area of western Guizhou: A case study of the Jiadi gold deposite in Panxian County [J]. Geology and Exploration, 56 (6): 1145-1157. (in Chinese with English abstract)

ZHANG K, YANG X K, YU H B, et al., 2020. Analysis of ore-controlling structure in the Changgou gold deposit of the northern Hanyin gold orefield, southern Qinling Mountains [J]. Journal of Geomechanics, 26 (3): 363-375. (in Chinese with English abstract)

ZHANG X Y, YE P, WANG S, et al., 2018. Geological entity recognition method based on Deep Belief Networks [J]. Acta

Petrologica Sinica，34（2）：343-351.（in Chinese with English abstract）

ZHU Y Q，TAN Y J，WU Y L，et al.，2017. Research on semantic retrieval model towards geological big data［J］. China Mining Magazine，26（12）：143-149.（in Chinese with English abstract）

ZHU Y Q，ZHOU W W，XU Y，et al.，2017b. Intelligent learning for knowledge graph towards geological data［J］. Scientific Programming，2017：5072427.

附中文参考文献

蔡强，郝佳云，曹健，等，2018. 采用多尺度注意力机制的远程监督关系抽取［J］. 中文信息学报，32（1）：96-101.

蔡强，李晶，郝佳云，2019. 基于聚焦损失与残差网络的远程监督关系抽取［J］. 计算机工程，45（12）：166-170.

陈建平，李靖，谢帅，等，2017. 中国地质大数据研究现状［J］. 地质学刊，2017，41（3）：353-366.

吕鹏飞，王春宁，朱月琴，2017. 基于文献的地质实体关系抽取方法研究［J］. 中国矿业，26（10）：167-172.

钱小梅，刘嘉勇，程芃森，2020. 基于密集连接卷积神经网络的远程监督关系抽取［J］. 计算机科学，47（2）：157-162.

宋明春，李三忠，伊丕厚，等，2014. 中国胶东焦家式金矿类型及其成矿理论［J］. 吉林大学学报（地球科学版），44（1）：87-104.

谭永杰，文敏，朱月琴，等，2017. 地质数据的大数据特性研究［J］. 中国矿业，26（9）：67-71，84.

唐朝，诺明花，胡岩，2020. ResNet 结合 BiGRU 的关系抽取混合模型［J］. 中文信息学报，34（2）：38-45.

汪青松，张金会，尤淼，等，2021. 井中矿层多要素探测方法研究与厚覆盖区金矿勘查试验：以皖东北五河地区为例［J］. 地质与勘探，57（1）：136-145.

薛玉山，王瑞廷，汪超，等，2020. 陕西省山阳县王家坪金矿断裂构造控矿规律［J］. 地质力学学报，26（3）：391-404.

张兵强，杨清毫，赵富远，等，2020. 贵州西部峨眉山玄武岩区金矿赋矿层位及矿石特征：以盘县架底金矿为例［J］. 地质与勘探，56（6）：1145-1157.

张康，杨兴科，于恒彬，等，2020. 南秦岭汉阴北部金矿田长沟金矿区控矿构造解析［J］. 地质力学学报，26（3）：363-375.

张雪英，叶鹏，王曙，等，2018. 基于深度信念网络的地质实体识别方法［J］. 岩石学报，34（2）：343-351.

朱月琴，谭永杰，吴永亮，等，2017. 面向地质大数据的语义检索模型研究［J］. 中国矿业，26（12）：143-149.

开放科学（资源服务）标识码（OSID）：

可扫码直接下载文章电子版，也有可能听到作者的
语音介绍及更多文章相关资讯