Computer Science and Application 计算机科学与应用, 2021, 11(11), 2672-2681
Published Online November 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1111271

Hans汉斯

地质灾害文本实体关系联合抽取研究

周榆婷[1]，陈平华[1]，陈建平[2]

[1]广东工业大学计算机学院，广东 广州
[2]肇庆学院计算机科学与软件学院，广东 肇庆

收稿日期：2021年10月5日；录用日期：2021年11月4日；发布日期：2021年11月12日

摘 要

针对传统实体关系联合抽取方法存在效率低下、错误传播、实体冗余等问题，提出基于双向长短时记忆神经网络和条件随机场并融合注意力机制的地质灾害实体与关系联合抽取方法。使用一种新标注方案，将地质灾害文本实体关系联合抽取问题转化为序列标注问题。用字符级嵌入进行作文本向量化表示，使用BiLSTM-Attention-CRF模型实现地质灾害文本实体关系联合抽取。实验结果表明：在地质灾害语料集上，实体识别的F-score值达到了85.4%，关系抽取的F-score达到了63.6%，证明了该方法的优越性和有效性。

关键词

命名实体识别，关系抽取，联合抽取，标注模式

Research on Joint Extraction of Entity and Relation in Geological Hazard Text

Yuting Zhou[1], Pinghua Chen[1], Jianping Chen[2]

[1]School of Computer, Guangdong University of Technology, Guangzhou Guangdong
[2]School of Computer Science and Software, Zhaoqing University, Zhaoqing Guangdong

Received: Oct. 5th, 2021; accepted: Nov. 4th, 2021; published: Nov. 12th, 2021

Abstract

In view of the problems of low efficiency, error propagation, and entity redundancy in traditional entities and relations extraction method, this article proposes a joint geological hazard entities and relations extraction method based on bi-directional long short-term memory and conditional random fields with attention mechanism. This method employs a new tagging scheme which

文章引用: 周榆婷, 陈平华, 陈建平. 地质灾害文本实体关系联合抽取研究[J]. 计算机科学与应用, 2021, 11(11): 2672-2681. DOI: 10.12677/csa.2021.1111271

represents both entity and relation information by the tags and converts the joint extraction task to a tagging task. This method applies character embedding as input, and extracts geological hazard entities and relations with BiLSTM-Attention-CRF model. The results shows that, on geological hazard corpus, the method achieves 85.4% F-score for named entity recognition and 63.6% F-score for relations extraction which proves the superiority and effectiveness of the method.

Keywords

Named Entity Recognition, Relation Extraction, Joint Extraction, Labeling Mode

1. 引言

近年来，日趋严重的地质灾害严重阻碍了地方经济发展和资源开发工作，给人民群众的生产生活造成困扰。科学家对地质灾害进行了深入研究与探索，随之产生了大量地质灾害文本。地质灾害文本中存在大量的知识三元组(实体－关系－实体)，使用文本挖掘技术自动抽取这些结构化信息，对地质灾害信息的统计与分析具有重要意义。其中，实体关系抽取已经在金融、医疗等领域取得了广泛的应用。从海量的非结构化地质灾害文本中抽取三元组，可以高效地管理及利用地质灾害信息，促进了地质灾害的预防和治理工作。

根据命名实体识别(Named Entity Recognition, NER)和关系抽取(Relation Extraction, RE)两个子任务完成的先后顺序，基于深度学习的实体关系抽取方法包括流水线方法和联合抽取方法。其中传统的流水线方法先完成 NER 子任务[1]，对识别到的实体进行两两匹配，再根据实体对进行 RE 子任务[2]。对于 NER 子任务，隐马尔可夫模型(Hidden Markov Model, HMM)和条件随机场(Conditional Random Field, CRF)等线性统计模型在该任务上表现出良好的性能[3]。随着深度学习的应用，基于神经网络的文本特征提取具有更好的识别性能，其中 BiLSTM-CRF 模型结合了循环神经网络(Recurrent Neural Network, RNN)和 CRF 模型的优点，在序列标注任务中表现出色[4]。但这种流水线的方法存在 2 个问题：1) 单独训练的 NER 模型和 RE 模型虽然提高了框架的灵活性，但忽略了两个子任务的相关性，难用利用潜在信息提高模型表达能力。2) NER 的错误会影响 RE 的结果，致使地质灾害知识三元组抽取效果较差。

与流水线方法不同，联合抽取方法旨在充分利用 NER 和 RE 的信息交互提升模型性能，目前基于联合抽取的三元组抽取研究已经在通用领域取得了广泛的应用。Miwa 等人[5]将实体关系抽取转化为表格填充问题，通过槽值填充的方式同时识别出句子的实体和关系，但该方法的性能取决于特征工程；为了避免复杂的特征工程，Miwa 等人[6]首次提出参数共享的方法，将循环神经网络应用在实体关系联合抽取任务中，在 ACE04、AEC05 数据集上的准确率显著提升；然而 Miwa 等人忽略了实体对之间长距离依赖的问题，为此，Zheng 等人[7]提出一种新的标注模式同时标注实体对和关系，将实体关系联合抽取任务看作序列标注问题，在新闻、服务等领域得到了广泛的应用。但该方法默认一个实体只能参与构建一个知识三元组，无法处理关系重叠问题。地质灾害文本中，一个实体通常以一对多的方式参与多种关系的构建，这样一对一的匹配规则会遗漏大量的三元组。为了促进 NER 和 RE 两个子任务的交互并识别重叠关系，本文提出一种改进的标注模式，将实体位置信息融入标签信息中，提出 BiLSTM 神经网络和 CRF

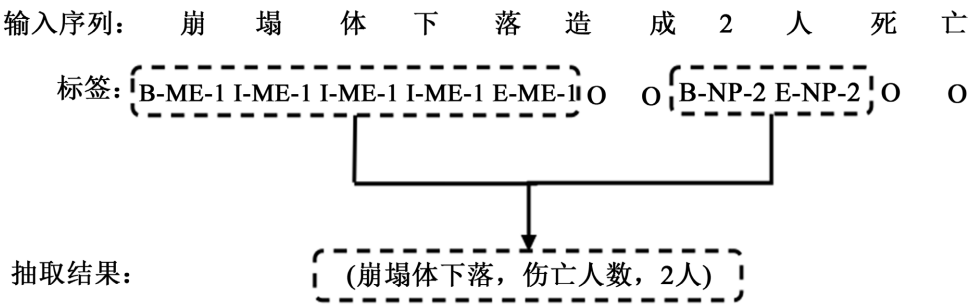
相结合的实体关系联合抽取方法，并融入注意力机制提升文本特征表达能力，使用 BiLSTM-Attention-CRF 模型将实体关系联合抽取任务转化为序列标注问题，本文主要有两方面的贡献：1) 针对 Zheng 等人[7]提出的标注模式在抽取重叠关系的不足，本文在标签信息中添加实体位置信息，优化了标注模式以及实体对匹配策略，显著提升了地质灾害文本中重叠关系抽取的准确率；2) 针对实体对之间距离长的问题，在端到端联合学习模型中加入注意力机制(Attention)，不仅能提取词级特征，还能加强句子级特征提取，有效缓解了长距离实体对之间的关系抽取问题。

本文实验在人工标注的地质灾害语料集上展开，相比于流水线方法，本文的实体关系联合抽取方法取得了更好的效果，地质灾害文本命名实体识别的 F-score 达到了 85.4%，关系抽取的 F-score 达到了 63.5%。

2. 标注方法与抽取规则

2.1. 标注方法

在一条数据仅围绕一个主实体(Main_Enity, ME)的地质灾害文本中进行实体和关系的联合抽取，本质上只需抽取与主实体 ME 存在关系的其他实体 $\{X_1, X_2, \cdots, X_i, \cdots, X_n\}$ 以及两个实体之间的关系类型 $\{R_1, R_2, \cdots, R_i, \cdots, R_n\}$，其中 X_i 表示与 ME 存在关系的第 i 个实体，R_i 表示 X_i 与 ME 之间的关系类型。本文基于 Zheng 等人[7]提出的标注模式，在标签信息中增加了实体位置信息，丰富了实体的标签含义，以处理重叠实体问题。本文提出的标签由实体边界、实体位置、关系类别三个部分构成。其中，实体边界标签由通用的命名实体标注方法“BIOES”表示，B 表示实体首字，I 表示实体中间位置，E 表示实体的结束位置，S 表示由单一字符构成的实体，O 表示非实体。关系类别是根据地质灾害语料集统计得到的，本文使用 3550 条地质灾害数据，共包含 12 种关系，分别为：时间、地点、气候、风速、气温、体积、降雨量、地层岩性、灾害点密度、伤亡人数、经济损失和防治措施。实体位置标签描述了实体在三元组中的位置，由 1 和 2 表示，其中 1 表示该实体位于三元组的第一个位置，2 表示第二个位置，最后实体关系联合抽取结果可以由三元组(实体 1，关系类别，实体 2)表示。以描述一次崩塌事故的数据为例，标注示例如图 1 所示。首先将“崩塌体下落”标注为固定标签主实体“ME”，在文本中，“崩塌体下落”与“2 人”之间存在关系“伤亡人数”，则将“2 人”标注为“伤亡人数”的代表标签“NP”(Number_of_People，NP)。此外，“崩塌体下落”是关系中的第一个实体，故增加位置标签“1”，“2 人”是关系中的第二个实体，增加位置标签“2”。当匹配到标签“ME”和“NP”的“BE”集合，即抽取到三元组(崩塌体下落，伤亡人数，2 人)。



注：ME为主实体，NP为伤亡人数。

Figure 1. Annotation example
图 1. 标注样例

本文标注方法只关注主实体与各实体间的关系类型而无需关注实体本身所属的实体类型，只在预定义关系集合上进行标注和抽取，减少无关实体对的错误转播和关系冗余。同时，对于 ME 和多个之间存在重叠关系的问题，增加实体位置信息明确实体在关系中的位置。此外，不同于传统标注方法和流水线方法割裂了 NER 和 RC 两个子任务的联系，本文方法对实体和关系进行同步标注，节省了标注成本。

2.2. 抽取规则

Zheng 等人[7]所遵循最近匹配原则默认一个实体只参与构建一个知识三元组，本文充分考虑地质灾害文本的特点，针对其中存在的大量重叠关系，改进了实体对匹配规则：

1) 根据实体边界信息从序列中取出实体，取实体第一个字的关系类别作为该实体的关系类别，同样将第一个字的位置信息作为该实体在三元组中的位置。

2) 按照最近匹配原则进行关系抽取，查找与该实体距离最近的实体，并判断两实体的关系类别和实体位置是否符合预定义的关系类型，若符合则构成一个知识三元组。

3) 关系类别为预定义的 12 种关系类别，实体在三元组中的位置由实体位置标签确定，位置标签为 1 的实体只能与位置标签为 2 的实体匹配。

4) 实体对匹配过程具有方向性，位置标签为 1 的实体只能向后匹配，位置标签为 2 的实体只能向前匹配，避免了无效查找和匹配。

以“崩塌体下落造成 2 人死亡，直接经济损失达 154.7 万元。”为例，伤亡人数代表标签为 NP，经济损失代表标签为 EL (Economic_Loss, EL)。首先根据规则 1 抽取到(崩塌体下落，ME，1)，(2 人，NP，2)和(154.7 万元，EL，2)三个实体。对于这三个实体，首先主实体“崩塌体下落”向后查找到实体“2 人”，由于两个实体的关系类别分别为 ME、NP，实体位置分别为 1、2，故符合规则，生成三元组(崩塌体下落，伤亡人数，2 人)；接着，实体“154.7 万元”向前查找到实体“2 人”，二者的实体的位置标签均为 2，故不匹配，继续向前查找；最后实体“154.7 万元”向前查找到实体“崩塌体下落”，二者的关系类别分别为 ME、EL，实体位置分别为 1、2，组成三元组(崩塌体下落，经济损失，154.7 万元)。

3. BiLSTM-Attention-CRF 模型

3.1. 网络总体结构

BiLSTM-Attention-CRF 模型由字嵌入层、BiLSTM 层、Attention 层和 CRF 层组成。模型的输入是按照字符级别切分的文本序列，该序列经过字嵌入层转换为字符向量，并作为 BILSTM 层的输入。字向量进入 BiLSTM 层，并引入注意力机制提取特征，得到包含上下文信息的文本序列双向表达，对其作拼接后，将拼接结果经过隐层映射后输入序列标注模型 CRF 层。CRF 层预测每个字符对应标签的概率，将其与真实标签作对比，运用动态优化泛得到最终标签。

网络结构如图 2 所示。

3.2. Embedding 层

模型的输入是按照字符级别切分的文本序列，Embedding 层的作用将文本序列表示为固定维度的向量，本文使用预训练的 Word2vec 词向量模型进行编码。图 2 可以看出，本文的 BiLSTM-Attention-CRF 模型基于字符向量。在地质灾害文本中，实体的表达通常具有相同的表达规范和度量单位，字符向量能够更好地表达地质灾害实体的结构特征。具体来说就是将句子中的每个字符用 Word2vec 词向量表示，最

后得到关于句子的向量表示序列，假设一个句子 X 有 n 个字，则该句的向量表达可表示为 $X=(x_1,x_2,\cdots,x_n)$，其中 $x_i \in R^d$，d 是字符 embedding 的维度。

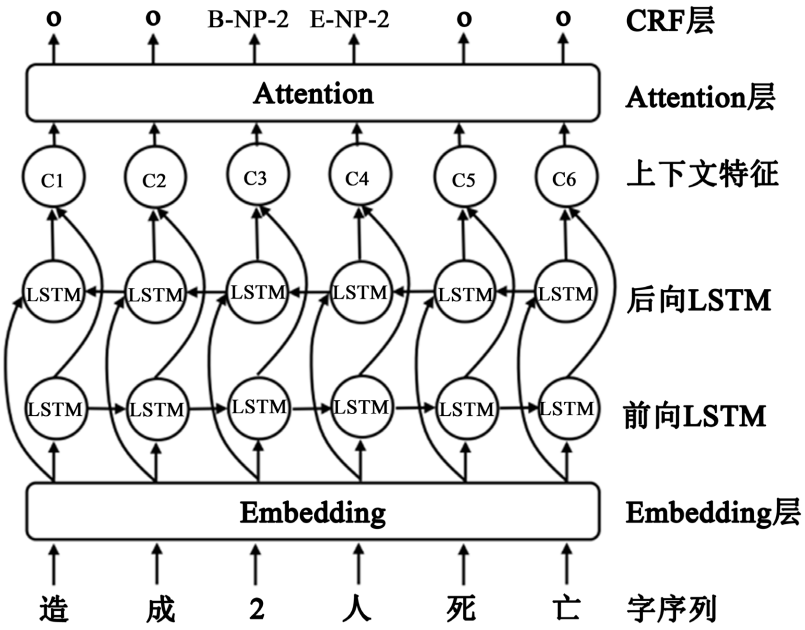


Figure 2. BiLSTM-Attention-CRF model
图 2. BiLSTM-Attention-CRF 模型

3.3. BiLSTM 层

循环神经网络具有短期记忆能力，自带时序性的特点使其适用于语音、视频等时序数据的建模[8]。本文使用的 LSTM 神经基于 RNN 作了较大改进，引入了门控机制来控制神经网络中信息的积累速度。LSTM 有输入门、输出门和遗忘门共 3 个门控单元。其中输入门控制当前时刻的信息输入到记忆细胞的比例，遗忘门决定了上一时刻的状态信息丢弃的比例，最终由输出门控制当前细胞状态的输出[9]。LSTM 使用门控机制缓解了 RNN 的长距离依赖和梯度消失问题，使得当前时刻的预测能够充分利用上下文语义特征[10] [11]。本文借鉴 Hong [12]等人使用的 LSTM 结构，其数学模型为：

$$i_t = \tanh\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i\right) \tag{1}$$

$$f_t = \tanh\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}C_{t-1} + b_f\right) \tag{2}$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{3}$$

$$o_t = \tanh\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}C_t + b_o\right) \tag{4}$$

$$h_t = o_t \tanh\left(c_t\right) \tag{5}$$

其中，W_* 是 LSTM 神经元的参数矩阵；b_* 是 LSTM 神经元的偏置项；使用 tanh 激活函数增加神经网络的非线性。LSTM 具有方向性，为了充分利用上下文信息，Lample 等人[13]提出双向长短时记忆神经网络(Bi-directional LSTM，BiLSTM)。前向 LSTM 考虑 x_t 的上文信息来学习上文特征 $h_{t,forward}$，后向 LSTM 考虑 x_t 的下文信息来学习下文特征 $h_{t,backward}$，再将二者拼接成 $\left[h_{t,forward}, h_{t,backward}\right]$，用于序列标注任务的文本特征提取。

3.4. Attention 层

模型的第三层是 Attention 层，主要作用是增强模型关注并学习重要信息的能力[14]。经过 BILSTM 编码的隐向量具有丰富的语义特征，然而在预测实体标签时这些特征的重要程度相同，导致较大的误差。

$$S = \sum \alpha_i \cdot x_i \tag{6}$$

Attention 机制将对得到的每个词向量 x_i，与权重 α_i，S 表示由 $x_1, x_2, \cdots, x_n$ 构成的句子。Attention 计算字符之间的相关性程度 α_i，便于识别实体边界，缓解了字符级分词造成的实体边界模糊问题。

3.5. CRF 层

与 HMM 不同，CRF 具有更宽松的条件独立性假设，特征设计更灵活，被广泛应用于序列标注任务中。本文使用的线性链条件随机场，是一种根据输入随机变量输出预测序列的条件概率分布模型[15] [16]。对于指定序列 $X = (x_1, x_2, \cdots, x_n)$，其对应标签 $Y = (y_1, y_2, \cdots, y_n)$，若满足下列条件：

$$P(Y_i \mid X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n) = P(Y_i \mid X, Y_{i-1}, Y_{i+1}) \tag{7}$$

设 P 为解码层输出的权重矩阵，维度为 $n \times k$，n 表示序列长度，k 表示标签集合的个数，从而可以得出评估分数 $S(x, y)$，即：

$$S(x, y) = \sum_{i=0}^{n} A_{yi, yi+1} + \sum_{i=0}^{n} P_{i, yi} \tag{8}$$

其中：x 为输入的文本序列，$A_{yi,yi+1}$ 为从标签 y_i 转移到 y_{i+1} 的分数，$A_{yi,yi+1}$ 的值越大表示标签 i 转移到标签 j 的可能性越大，$P_{i,yi}$ 为第 i 个字被预测为第 y_i 个标签的分数。

4. 实验

4.1. 实验设置

本文实验脚本语言使用 python3.7，基于 pytorch 1.5.0，运行在 Intel Core i5、1.8GHz CPU、8 GB RAM、macOS 10.13 操作系统等软硬件环境。

本文实验的数据来自中国地质调查局地质环境监测院公开的地灾防治文档，及百度爬取的危险性评估、应急调查报告，从文档中抽取 3550 条地质灾害数据进行实验，该数据集标注了 5 种类型的地质灾害，包括滑坡、崩塌、泥石流、地面塌陷、地裂缝，语料集信息如表 1 所示。

Table 1. Geological hazard corpus information
表 1. 地质灾害语料集信息

数据类型	文档数量	训练集	测试集
滑坡	861	689	172
崩塌	850	680	170
泥石流	773	618	155
地面塌陷	712	570	142
地裂缝	354	283	71

本文随机抽取 10%的训练样本作为验证集，用于调整神经网络的超参数，本文模型的超参数设置如表 2 所示：

Table 2. Model hyperparameters
表 2. 模型的超参数

参数	设置
字向量维度	100
BiLSTM 网络神经元数量	200
BiLSTM 网络层数	1
学习率	0.01
Batch_size	64
Dropout	0.5

本文采用信息抽取领域的三项基本指标，准确率(Precision, P)、召回率(recall, R)和 F 值(F-score, F)对预测结果进行评价：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} \times \text{Recall}} \tag{11}$$

4.2. 标注模式对比实验

本节使用基于字符嵌入的 BiLSTM-Attention-CRF 模型，以 Zheng 等人[7]提出的标注模式和最近距离匹配原则作为 baseline，依次引入本文改进的抽取规则、增加位置标签，对比了不同的标注模式和抽取规则的性能，实验结果如表 3 所示，可以得到 2 个结论：

Zheng 等人[7]所遵循最近匹配原则默认一个实体只参与构建一个知识三元组，在地质灾害文本上会损失大量的重叠关系。使用本文改进的抽取规则关系抽取的 F-score 提高了 12.1%，其中重叠关系取得了 27.3%的召回率提升，说明了本文改进的抽取规则可以更充分地应对实体之间一对多匹配的情况。

基于本文的抽取规则，增加位置信息标签后，关系抽取的 F-score 值提升了 9.3%，其中重叠关系取得了 9.1%的召回率提升。说明增加位置信息来表示实体在三元组中的位置是可行的，同时表明神经网络可以学习到位置标签。

Table 3. Performance comparison of annotation modes
表 3. 标注模式的性能比较

方法	RE			RE (Overlapping)
	P	R	F	R
Baseline	0.651	0.311	0.421	0.153
+本文提取规则	0.632	0.474	0.542	0.426
+位置标签	0.822	0.517	0.635	0.517

4.3. 模型对比实验

本节使用字符向量作为输入特征，选取了流水线方法中的 BiLSTM-CNN 模型和联合学习方法中的

LSTM-CRF、BiLSTM-CRF 和 BiLSTM-Attention-CRF 模型进行对比试验。其中，基于流水线的方法采用传统的“BIOES”方法标注实体，使用 BiLSTM 搭建实体抽取模型，匹配实体对，再使用 CNN 搭建关系抽取模型。基于联合学习的实体和关系抽取方法，采用本文改进的标注模式及抽取规则，设置 LSTM-CRF、BiLSTM-CRF 和 BiLSTM-Attention-CRF 端到端序列标注模型进行对比实验，实验结果如表 4 所示：

Table 4. Performance comparison of models
表 4. 模型的性能比较

模型	命名实体识别			关系抽取		
	P	R	F	P	R	F
BiLSTM-CNN	0.849	0.854	0.851	0.807	0.225	0.352
LSTM-CRF	0.863	0.871	0.867	0.745	0.454	0.564
BiLSTM-CRF	0.856	0.883	0.869	0.784	0.474	0.591
BiLSTM-Attention-CRF	0.878	0.832	0.854	0.822	0.517	0.635

由表 4 可以得到 2 个结论：

1) BiLSTM-Attention-CRF 模型在 NER 任务上的 F 值为 85.4%，关系抽取子任务的 F 值为 63.5%，优于其他较先进模型在此任务上的表现，证实了了本文算法的有效性。此外从表 4 还可得出，在地质灾害文本上，本文的抽取规则和标注模式明显优于流水线方法。证实了本文的标注模式应用于地质灾害文本的实体关系联合抽取是有效的。

2) 从准确率数据来看，与 LSTM-CRF 模型相比，使用 BiLSTM 进行编码的模型的准确率有了显著提升，原因可能是 BiLSTM 能充分利用上下文信息，捕捉长距离内容，提升了对文本的处理与表示能力。在所有联合抽取模型中，只有 BiLSTM-Attention-CRF 模型能获得更高的准确率和召回率，得到最高的 F 值，说明添加注意力机制的网络结构能够有效识别实体边界，一定程度上缓解了边界模糊的问题，对地质灾害文本的特征表示能力更好。

4.4. 模型预测结果分析

基于本文的标注策略，BiLSTM-Attention-CRF 模型对主实体与各实体间关系的预测结果如表 5 所示，整体效果较为均衡。

Table 5. Prediction results of BiLSTM-Attention-CRF model on entities
表 5. BiLSTM-Attention-CRF 模型对实体的预测结果

主实体及关系类型	P	R	F1
主实体	0.951	0.892	0.921
时间	0.904	0.865	0.884
地点	0.895	0.792	0.840
气候	0.925	0.907	0.916
风速	0.912	0.834	0.871
气温	0.903	0.869	0.886
塌方量	0.873	0.833	0.853

Continued

降雨量	0.863	0.844	0.853
地层岩性	0.792	0.723	0.756
灾害点密度	0.832	0.796	0.814
伤亡人数	0.891	0.879	0.885
经济损失	0.867	0.841	0.854
防治措施	0.811	0.744	0.776

其中，地质灾害主实体、时间、风速、气候和气温具有较高的准确率和 F 值，均高于 90%，但地层岩性和防治措施实体预测结果明显低于平均水平，召回率分别为 75.6%和 77.6%。

分析文本语料得出，风速、气候等实体具有较规则的计量单位或后缀组成词，如风速的“m/s”、气候的“季风气候”等，这些明显的字特征信息提高了实体识别的准确率。而地层岩性的构词比较复杂，如“颜色 + 岩性”、“颜色 + 形状 + 粒度 + 岩性”等方式，引入大量噪声。防治措施的召回率低是由于地质灾害防治措施的多样化、文本长度相差较大、描述方法不统一，实体边界模糊，如“生物措施”、“采用强夯或分层碾压措施降低土的压缩性”、“清理地基顶面草根、树根和各种杂物”等描述。由此可见，边界错误是地层岩性和防治措施召回率偏低的主要因素，本文采用了较严格的判定方式，即实体边界和位置信息需要同时预测正确，对于一部分实体，尽管预测边界和标注序列存在差异，但也具有一定的意义，在实际应用中仍然可以被使用，例如在上述描述防治措施的例子中，若模型识别结果为“强夯或分层碾压措施”或“降低土的压缩性”也具有实用价值。

此外，本文数据集由人工标注获得、基于中文分词、依存分析等自然语言处理工具输出的结果，这些预处理方法所产生的误差会直接累积到实体关系抽取阶段。

5. 结语

本文面向地质灾害文本知识三元组抽取任务，标注了地质灾害文本，弥补了该领域上语料集的缺失。改进了关系抽取规则和标注模式，增加了位置标签，相比最近匹配原则，本文的方法能够显著提升重叠关系的召回率。使用 BiLSTM-Attention-CRF 模型，将地质灾害文本实体关系联合抽取转化为端到端的序列标注任务，注意力机制有效提高了边界识别的准确率。实验表明，本文所提方法在地质灾害文本上，NER 的 F-score 达到了 85.4%，关系抽取的 F-score 值达到了 63.6%。由于关系抽取的准确率较低，下一步工作将研究深度学习与规则相结合的三元组抽取方法。

基金项目

广东省科技计划项目(2020B1010010010，2019B101001021)；

广东省自然科学基金项目(2019A1515010700)。

参考文献

[1] Zhang, R., Lu, W., Wang, S., Peng, X., Yu, R. and Gao, Y. (2020) Chinese Clinical Named Entity Recognition Based on Stacked Neural Network. *Concurrency and Computation*: *Practice and Experience*, **33**, e5775. https://doi.org/10.1002/cpe.5775

[2] 何玉洁, 杜方, 史英杰, 宋丽娟. 基于深度学习的命名实体识别研究综述[J/OL]. 计算机工程与应用, 2021: 1-17. http://kns.cnki.net/kcms/detail/11.2127.TP.20210326.0937.002.html, 2021-04-13.

[3] 西尔艾力·色提, 艾山·吾买尔, 王路路, 等. 结合单词-字符引导注意力网络的中文旅游文本命名实体识别[J]. 计

算机工程, 2021, 47(2): 39-45.

[4] Gao, S., Kotevska, O., Sorokine, A. and Christian, J.B. (2021) A Pre-Training and Self-Training Approach for Biomedical Named Entity Recognition. *PLoS ONE*, **16**, e0246310. https://doi.org/10.1371/journal.pone.0246310

[5] Miwa, M. and Sasaki, Y. (2014) Modeling Joint Entity and Relation Extraction with Table Representation. *Proc of the* 19*th Conf on Empirical Methods in Natural Language Processing*, Doha, October 2014, 1858-1869. https://doi.org/10.3115/v1/D14-1200

[6] Miwa, M. and Bausal, M. (2016) End to-End Relation Extraction Using LSTMs on Sequences and Tree Structures. *Proc of the* 54*th Association for Computational Linguistics*, Berlin, August 2016, 1105-1116. https://doi.org/10.18653/v1/P16-1105

[7] Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P. and Xu, B. (2017) Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Vancouver, July 2017, 1227-1236. https://doi.org/10.18653/v1/P17-1113

[8] Lee, S.H., Goëau, H., Bonnet, P. and Joly, A. (2020) Attention-Based Recurrent Neural Network for Plant Disease Classification. *Frontiers in Plant Science*, **11**, Article ID: 601250. https://doi.org/10.3389/fpls.2020.601250

[9] 张华丽, 康晓东, 李博, 王亚鸽, 刘汉卿, 白放. 结合注意力机制的 Bi-LSTM-CRF 中文电子病历命名实体识别[J]. 计算机应用, 2020, 40(S1): 98-102.

[10] Deng, N., Fu, H. and Chen, X. (2021) Named Entity Recognition of Traditional Chinese Medicine Patents Based on BiLSTM-CRF. *Wireless Communications and Mobile Computing*, **2021**, Article ID: 6696205. https://doi.org/10.1155/2021/6696205

[11] Kadyan, V., Dua, M. and Dhiman, P. (2021) Enhancing Accuracy of Long Contextual Dependencies for Punjabi Speech Recognition System Using Deep LSTM. *International Journal of Speech Technology*, **24**, 517-527. https://doi.org/10.1007/s10772-021-09814-2

[12] Hong, Y., Liu, Y., Yang, S., Zhang, K. and Hu, J. (2020) Joint Extraction of Entities and Relations Using Graph Convolution over Pruned Dependency Trees. *Neurocomputing*, **411**, 302-312. https://doi.org/10.1016/j.neucom.2020.06.061

[13] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, Chris (2016) Neural Architectures for Named Entity Recognition. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2019, 5998-6008.

[15] Wojek, C. and Schiele, B. (2008) A Dynamic Conditional Random Field Model for Joint Labeling of Object and Scene Classes. *European Conference on Computer Vision*, Marseille, 12-18 October 2008, 733-747. https://doi.org/10.1007/978-3-540-88693-8_54

[16] Precheh, I. (1998) Automatic Early Stopping Using Cross Validation: Quantifying the Criteria. *Neural Networks the Official Journal of the International Neural Network Society*, **11**, 761-767. https://doi.org/10.1016/S0893-6080(98)00010-0